2013 IEEE 13th International Conference on Data Mining

Linear Computation for Independent Social Influence

Qi Liu[1], Biao Xiang[1], Lei Zhang[1], Enhong Chen[1,*], Chang Tan[1], Ji Chen[2]
[1]*School of Computer Science and Technology, University of Science and Technology of China*
{qiliuql,cheneh}@ustc.edu.cn, {bxiang,stone,tanchang}@mail.ustc.edu.cn
[2]*Microsoft, One Microsoft Way, Redmond, WA 98052, USA*
jich@microsoft.com

***Abstract*—Recent years have witnessed the increased interests in exploiting influence in social networks for many applications. To the best of our knowledge, from the computational aspect of social influence analysis, most of existing work focus on either describing the influence propagation process or identifying the set of most influential seed nodes. However, these work usually do not distinguish the "independent influence" of each single seed node after removing other seeds. Since it is important to quickly figure out the real contribution of each seed, in this paper we propose to measure the seed's independent influence by a linear social influence model. Specifically, we first describe the linear social influence model, and then define the independent influence under this model for eliminating the "mutual enrichment" between seed nodes. Meanwhile, we find that the influence of a set of nodes is actually the sum of their independent influence, and we also give upper bounds for independent influence. Moreover, these findings are evaluated by two applications, i.e., ranking the seeds by their independent influence and identifying the Top-K influential ones. Finally, the experimental results on several real-world datasets validate the effectiveness and efficiency of the proposed independent social influence measures.**

I. Introduction

Social networks have become very popular and they provide unparallel opportunities for understanding the human world and building novel applications. Thus, there is much research on social network analysis [1]. Indeed, tremendous efforts have been made to analyze and exploit the social influence between individuals for marketing, advertisement, recommendations, and so on [2].

Social influence, as an intuitive and well-accepted phenomenon, refers to the behavioral change (e.g., opinions, decisions) of individuals (nodes) affected by others in a network. From the computational aspect (i.e., qualitatively or quantitatively measuring the social influence), two of the major research interests are modelling influence propagation and locating the set of most influential seed nodes by social influence computation. Actually, several models [3], [4], [5] have been provided to describe the dynamics of influence propagation, and among them, Independent Cascade (IC) model [4] is one of the most popular and widely used models. However, it is usually hard to directly apply traditional influence models, like IC model, for social influence computation, since they are step-by-step algorithms and thus are very time consuming for large-scale social networks. Considering that we are often interested in finding the set of most influential nodes (called as a set of seeds) for social influence maximization, several greedy (e.g., CELF [6]) and heuristic (e.g., DegreeDiscount [7], PMIA [8] and IRIE [9]) approaches based on existing influence models have been proposed to improve the computational efficiency of the social influence maximization. Generally, existing approaches first claim an influence model, and then aim at finding a set of seed nodes by measuring their total influence, and finally use the seeds with a maximal influence for further applications, e.g., viral marketing [8].

However, to the best of our knowledge, few of scholars pay attention to the problem of efficiently measuring the independent social influence of these selected seed nodes. Here, we define the independent social influence of one node as its influence after removing others from the seed set (this seed set may be selected manually or by above algorithms, e.g., CELF). Intuitively, it is very important to figure out the real contribution (independent influence) of each seed node. For instance, the system in viral marketing could pay for the seeds based on their independent influence, or remove the less influential seeds from the seed set according to the budget. To that end, we propose to quickly measure each seed's independent influence by a linear social influence model that was preliminarily introduced in [10]. Specifically, we first describe the linear model, and then define the independent influence for eliminating the "mutual enrichment" between seed nodes. Meanwhile, we find two properties of the independent influence under this definition. The first property is that the influence of a set of nodes is actually the sum of their independent influence. The second property is that the independent influence of each single seed is no bigger than its original influence. Moreover, we evaluate these properties by two applications, i.e., ranking the seeds by their independent influence and identifying the Top-K influential seeds from the seed set. Finally, the experimental results validate the effectiveness and efficiency of the proposed independent social influence measures. Our main contributions can be summarized as follows.

- We propose the idea of efficiently measuring the independent social influence of each selected seed. Along this line, we formalize this computation problem under a linear social influence model.

*Contact Author.

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.48

- We provide two properties for the defined independent influence: the influence of a set of nodes is the sum of their independent influence; each node's original influence is the upper bound of its independent influence.
- The extensive experiments demonstrate that linear computation could efficiently and effectively rank the seeds by their independent influence, and the upper bounds could be used for quickly finding the nodes with the highest independent influence from the seed set.

The rest of this paper is organized as follows. Section II introduces the related work. In Section III, we present the linear social influence model and the measurement of independent social influence in detail. Section IV first demonstrates the two properties of the independent influence, and then applies independent influence for two applications. In Section V, we show the experimental results. Finally, Section VI concludes this paper.

II. Related Work

Several different kinds of research issues have been proposed in the context of social influence analysis [2]. The first issue is the measurement of the influence between two neighbor nodes in a social network. For instance, Anagnostopoulos et al. [11] differentiate social influence from homophily or confounding variables by proposing the shuffle test and edge reversal test. Goyal et al. [12] propose a model to learn the probabilities on social edges from a log of actions by the users. Moreover, Steeg et al. [13] introduce content transfer, an information-theoretic measure with a predictive interpretation to directly quantify the strength of the influence effect of one social user's content on another's.

The second and a central problem is to describe the dynamics of influence propagation in social networks [3], [4], [5], [14], [8], [15]. Among existing models, the idea of Independent Cascade (IC) model [4] and Linear Threshold (LT) model [5] are widely used. For instance, in IC model, the activated/influenced nodes have a single chance to influence their neighbors independently with a probability. This iterative propagation process will not stop until there is no newly influenced node. The IC model where each link shares the same propagation probability is called the Uniform IC Model, and the one with edge weights is called the Weighted Cascade (WC) Model [16]. Researchers have proved that the influence spread (i.e., the expected number of nodes that will be influenced) computation under IC model is #P-hard [8]. As an alternative, Monte-Carlo simulation, which is very-time-consuming, is employed to approximately calculate influence. Recently, Yang et al. propose GS ($Gauss-Seidel$) algorithm for quick approximation of influence spread under IC model by solving a linear system [17].

The third research goal is to apply social influence and social influence models to the real applications. For instance, by exploiting social influence, Li et al. propose IPRank algorithm for ranking both individuals and groups [18], and Yang et al. design a recommendation to provide guidance for a social user to systematically approach his friending target [19]. Actually, one of the most important application for social influence is viral marketing, and the research problem can be summarized as finding a set of seed nodes which will influence the maximal number of individuals in the social network for maximizing the word-of-mouth propagation of one product. Along this line, both greedy and heuristic approaches have been proposed. For instance, for finding a set of nodes for social influence maximization, Leskovec et al. design the cost-effective lazy forward (CELF) optimization [6] by taking advantage of submodularity property to reduce the number of searched nodes, and Chen et al. propose both the Degree Discount heuristic and the Maximum Influence Path heuristic exploiting the local structures of each node [7], [8]. However, to the best of our knowledge, the problem of measuring the independent social influence of each selected seed remains pretty much open.

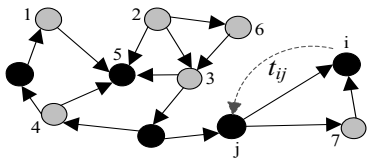


Figure 1. An example directed network.

III. Independent Social Influence

In this section, we first introduce the linear social influence model [10] which is both tractable and efficient. Then, we define the independent influence under this model.

Assume that $G = (V, A, \mathbf{T})$ is a network (as shown in Figure 1), where $V = \{1, 2, ..., n\}$ is the node set and edge set A represents all the connections between nodes. $\mathbf{T} = [t_{ij}]_{n*n}$ is a transmission matrix for influence propagation, t_{ij} represents the propagation probability from node i to node j. If there is an edge from j to i in A (i.e., j trusts i), then $t_{ij} > 0$[1], otherwise $t_{ij} = 0$. Since learning the non-zero t_{ij} [12] is beyond the scope of this paper, we assume they are known and usually $\sum_{i=1}^{n} t_{ij} \leq 1$ [17]. Here, G is assumed to be directed, as influence propagation is specific to direction in the most general case [14][2]. For better illustration, Table I shows some other math notations used in this paper.

A. Linear Social Influence Model

Formulation. In the literature of influence propagation, there are two well-known assumptions [4], [5]: (1) if one is the initiator of something (e.g. opinion), he/she will spread that with 100% probability; (2) otherwise, this probability will depend on his/her neighbors' influence. Following them, we could propose an influence model as:

[1]If j trusts i, then i influence j.

[2]The proposed techniques can be applied to undirected networks.

Table I
SEVERAL IMPORTANT MATHEMATICAL NOTATIONS.

Notations	Description
$f_{i\to j}$	influence from node i to j
$f_{i\to T}$	total influence from i to the nodes in set T
$\mathbf{f}_i$	influence vector for node i
d_j	parameter, the damping coefficient of node j
$\boldsymbol{\nu}_\mathbf{i}$	vector, $\nu_{\mathbf{i},i}$ is used to guarantee $f_{i\to i}=1$
$\mathbf{P}$	represents $(\mathbf{I}-d\mathbf{T}')^{-1}$, with each entry p_{ij}, each column $\mathbf{P}_{\cdot i}$
$\boldsymbol{p}$	vector, where $p_i=\sum_{k=1}^{n} p_{ki}$
$S\backslash\{i\}$	the nodes in S where node i is excluded
$f_{i\to j}^{S\backslash\{i\}}$	influence from node i to j (independent from other nodes in S)
$f_{i\to T}^{S\backslash\{i\}}$	independent influence from i to the nodes in set T
$\mathbf{f}_i^{S\backslash\{i\}}$	independent influence vector for node i
$\boldsymbol{\nu}_\mathbf{i}^{S\backslash\{i\}}$	vector, $i\in S$ and $\nu_{\mathbf{i},i}^{S\backslash\{i\}}$ is used to guarantee $f_{i\to i}^{S\backslash\{i\}}=1$
$\mathbf{T}_{\overline{S^iS^i}}$	matrix, removing the rows and columns corresponding to the members in $S\backslash\{i\}$ from $\mathbf{T}$
Γ	entry Γ_{ij}, each column $\Gamma_{\cdot i}$

Definition 1: Denote the influence from node i to j by $f_{i\to j}$, then

$$f_{i\to i} = 1, \tag{1}$$

$$f_{i\to j} = d_j\sum_{k\in N_j} t_{kj}f_{i\to k}, \quad for\ j\neq i, \tag{2}$$

where $N_j=\{j_1,j_2,...j_m\}$ is j's trust-friends set (i.e., $\forall k\in N_j$, there is a connection $(j,k)\in A$). The major difference of this definition from the traditional models is that we assume the influence flowing to node j is proportional to the linear combination of the influence to j's neighbors (see Equation (2)). Thus, the computation of influence will be in a linear efficient way. Here, the parameter d_j is the damping coefficient of j for the influence propagation[3]. It locates in range $[0,1]$, and the smaller d_j is, the more influence will be blocked by node j. For simplicity, we choose the same d for each node. Similarly, we denote $f_{i\to T}=\sum_{j\in T} f_{i\to j}$ as the influence spread from node i to a set of nodes T; that is, it stands for the total influence to the entire network if $T=V$.

Influence Computation. Under the above model definition, we can solve the influence spread vector $\mathbf{f}_i=[f_{i\to 1},f_{i\to 2},...f_{i\to n}]'$ for each node i as follows. First, we can rewrite Equation (1) and Equation (2) as

$$\mathbf{f}_i = d\mathbf{T}'\mathbf{f}_i+\boldsymbol{\nu}_\mathbf{i}=(\mathbf{I}-d\mathbf{T}')^{-1}\boldsymbol{\nu}_\mathbf{i} \tag{3}$$

$$= \mathbf{P}\boldsymbol{\nu}_\mathbf{i}, \tag{4}$$

where $\boldsymbol{\nu}_\mathbf{i}=[0,0,...,\nu_{\mathbf{i},i},...0]'$ is a vector with only the i-th entry $\nu_{\mathbf{i},i}$ nonzero; that is, $\nu_{\mathbf{i},i}$ should be equal to a number to guarantee $f_{i\to i}=1$ as described in Equation (1). In this equation, $(\mathbf{I}-d\mathbf{T}')$ is invertible because its transpose is strictly diagonally dominant, and $n*n$ matrix $\mathbf{P}=(\mathbf{I}-d\mathbf{T}')^{-1}$. As $\boldsymbol{\nu}_\mathbf{i}$ is a vector with only $\nu_{\mathbf{i},i}$ nonzero, Equation (4) could be rewritten as $\mathbf{f}_i=\nu_{\mathbf{i},i}\mathbf{P}_{\cdot i}$. Specifically, $f_{i\to i}=\nu_{\mathbf{i},i}p_{ii}$, with Equation (1), we could get

$$\nu_{\mathbf{i},i}=\frac{1}{p_{ii}}, \quad and\ thus, \quad \mathbf{f}_i=\frac{\mathbf{P}_{\cdot i}}{p_{ii}}. \tag{5}$$

[3]Which is similar to that in PageRank [20].

Since $\mathbf{P}$ is a positive definite matrix, $p_{ii}>0$. Then, the total influence from node i to the entire network G should be

$$f_{i\to V}=\mathbf{f}_i'\mathbf{e}=\sum_{j=1}^{n} f_{i\to j}=\frac{1}{p_{ii}}\sum_{j=1}^{n} p_{ji}=\frac{p_i}{p_{ii}}, \tag{6}$$

where $p_i=(\mathbf{P}_{\cdot i})'\mathbf{e}=\sum_{j=1}^{n} p_{ji}$. Given parameter d, and the influence transmission matrix $\mathbf{T}$, to get the influence vector $\mathbf{f}_i$, we only need to compute the i-th column of $\mathbf{P}$ ($\mathbf{P}_{\cdot i}$), which can be computed in $O(|A|)$ since $\mathbf{P}^{-1}\mathbf{P}_{\cdot i}=\mathbf{e}_i$ is a linear system and it satisfies the Gauss-Seidel condition. The computation process is shown in Algorithm 1, where the i-th entry in $\mathbf{e}_i$ ($\mathbf{e}_{i,i}$) is 1, otherwise, 0.

Algorithm 1: Gauss-Seidel: $(\mathbf{I}-d\mathbf{T}')\mathbf{P}_{\cdot i}=\mathbf{e}_i$ for $\mathbf{P}_{\cdot i}$

```
input : T, d, i
output: P_{·i}=[p_{1i}, ..., p_{ni}]': the i-th column of P.
for (j=0;j<n;j++) do
    p_{ji}^{(0)} = 0; //Initialization
iter=0;
while NOT-Converge do
    for (j=0;j<n;j++) do
        p_{ji}^{(iter+1)} = (e_{i,j} + Σ_{k=1}^{j-1} dt_{kj} p_{ki}^{(iter+1)} + Σ_{k=j+1}^{n} dt_{kj} p_{ki}^{(iter)});
    iter++;
return P_{·i}
```

Relationships with Traditional Models. Actually, in this paper, we use a specialization of the linear influence model proposed in [10] with the prior probability $\alpha_i=1$ for each node, and this linear model has close relationship with the traditional ones. For instance, it is easy to prove that the linear approximation method for the IC model [17] is actually a specialization of our linear model when $d=1$. Also, the non-linear stochastic model [14] can be well approximated by this model when $d\in(\frac{1}{2},1]$. Finally, it is worth noting that both PageRank and our model are random walk based methods, while in PageRank $f_{i\to i}=p_{ii}$ (a value for quick computation) rather than 1, and the detailed proof and explanations could be found in [10].

B. Independent Social Influence Computation

Definition. One drawback when applying the influence model illustrated in Section III-A is that it does not consider the "mutual enrichment" and "influence overlap" of different nodes. For instance, in a scientific collaboration network, if node (researcher) i is a close collaborator of j, and j is one of the most influential researcher in the network. Though i has limited influence herself, with the help of j (i.e., some of j's influence will flow to i), the computed influence of i will be much higher than its real value.

However, it seems hard to compute the real independent influence, since influence is always spread with the help of

others. Even though, given a node set S and a node $i \in S$, it is still possible to evaluate the independent influence $f_{i\to j}^{S\setminus\{i\}}$ from node i to j, independent from other nodes in S, i.e., $S\setminus\{i\}$. Based on the linear model, it could be defined as:

Definition 2: Denote the independent influence from node i to j (independent from other nodes in S, and $i \in S$) by $f_{i\to j}^{S\setminus\{i\}}$, then

$$f_{i\to i}^{S\setminus\{i\}} = 1, \tag{7}$$
$$f_{i\to j}^{S\setminus\{i\}} = 0, \quad j \in S \,\&\, j \neq i, \tag{8}$$
$$f_{i\to j}^{S\setminus\{i\}} = d\sum_{k\notin S} t_{kj} f_{i\to k}^{S\setminus\{i\}}, \quad j \notin S. \tag{9}$$

From this definition we can see that $f_{i\to j}^{S\setminus\{i\}}$ is essentially the influence of i when other nodes of S are "removed" from the network, i.e., these nodes stop receiving and forwarding the information from i. Similarly, the independent influence spread from node i to a set of nodes T could be denoted as $f_{i\to T}^{S\setminus\{i\}} = \sum_{j\in T} f_{i\to j}^{S\setminus\{i\}}$.

Independent Influence Computation. Similar to the influence computation illustrated in Section III-A, the independent influence spread vector $\mathbf{f}_i^{S\setminus\{i\}} = [f_{i\to 1}^{S\setminus\{i\}}, f_{i\to 2}^{S\setminus\{i\}}, ...f_{i\to n}^{S\setminus\{i\}}]'$ for node i could be computed as follows. First, we rewrite the equations in Definition 2 and have

$$\mathbf{f}_i^{S\setminus\{i\}} = (\mathbf{I} - d\mathbf{T}'_{\overline{S^i S^i}})^{-1}\nu_{\mathbf{i}}^{S\setminus\{i\}}, \tag{10}$$

[4]where $\mathbf{T}_{\overline{S^i S^i}}$ is the matrix reduced from $\mathbf{T}$ by removing it's rows and columns corresponding to the members in $S\setminus\{i\}$, and $\nu_{\mathbf{i}}^{S\setminus\{i\}} = [0, 0, ..., \nu_{\mathbf{i},i}^{S\setminus\{i\}}, ...0]'$ is a vector with only the i-th entry $\nu_{\mathbf{i},i}^{S\setminus\{i\}}$ nonzero, i.e., $\nu_{\mathbf{i},i}^{S\setminus\{i\}}$ should be equal to a number to guarantee $f_{i\to i}^{S\setminus\{i\}} = 1$. We could find Equation (10) is similar to Equation (3), and thus independent influence can be solved in the same way as shown in Section III-A. Specifically, if we denote $\Gamma = (\mathbf{I} - d\mathbf{T}'_{\overline{S^i S^i}})^{-1}$, then

$$\nu_{\mathbf{i},i}^{S\setminus\{i\}} = \frac{1}{\Gamma_{ii}}, \quad \textit{and thus}, \quad \mathbf{f}_i^{S\setminus\{i\}} = \frac{\Gamma_{\cdot i}}{\Gamma_{ii}}, \tag{11}$$

$$f_{i\to V}^{S\setminus\{i\}} = (\mathbf{f}_i^{S\setminus\{i\}})'\mathbf{e} = \sum_{j\in V-S} f_{i\to j}^{S\setminus\{i\}} = \frac{1}{\Gamma_{ii}}\sum_{j\in V-S}\Gamma_{ji}. \tag{12}$$

IV. Properties and Applications

In this section, we first demonstrate two properties of the proposed independent influence. With the help of these properties, we then apply independent influence to two possible applications.

A. Properties

Total Influence and Independent Influence. Following the definition in Section III-A, total influence spread from a node set S[5] to the network is $f_{S\to V} = \sum_{k\in V} f_{S\to k}$, and the influence spread vector $\mathbf{f}_S = [f_{S\to 1}, f_{S\to 2}, ...f_{S\to n}]'$ could be solved by

$$\mathbf{f}_S = d\mathbf{T}'\mathbf{f}_S + \nu_{\mathbf{S}} = (\mathbf{I} - d\mathbf{T}')^{-1}\nu_{\mathbf{S}} \tag{13}$$
$$= \mathbf{P}\nu_{\mathbf{S}}, \tag{14}$$

where $\nu_{\mathbf{S}} = [0, 0, ..., \nu_{\mathbf{S},i}, ...0]'$ is a vector with only the entries $\nu_{\mathbf{S},i}$ ($i \in S$) nonzero; that is, $\nu_{\mathbf{S},i}$ should be equal to a number to guarantee $f_{S\to i} = 1$. Equation (14) could be solved the same as Equation (4). In this way, a single influence value $f_{S\to V}$ is output, but it can not distinguish the contribution from each single node in S.

Actually, by the following theorem, we could find that this total influence ($f_{S\to V}$) is the sum of each single node's independent influence ($\sum_{i\in S} f_{i\to V}^{S\setminus\{i\}}$).

Theorem 1: For $\forall k \in V$, $f_{S\to k} = \sum_{i\in S} f_{i\to k}^{S\setminus\{i\}}$ and thus $f_{S\to V} = \sum_{i\in S} f_{i\to V}^{S\setminus\{i\}}$.

Proof: First, we define an auxiliary function as $\mathbf{g}(\mathbf{a}) = [g_1(\mathbf{a}), g_2(\mathbf{a}), ...g_n(\mathbf{a})]'$, where $\mathbf{a} = [a_1, a_2, ..., a_{|S|}]'$ and

$$g_i(\mathbf{a}) = a_i, \quad i \in S, \tag{15}$$
$$g_j(\mathbf{a}) = d\sum_{k=1} t_{kj} g_k(\mathbf{a}), \quad j \notin S, \tag{16}$$

From Appendix 1, we could prove that $\mathbf{g}(\mathbf{a}+\mathbf{b}) = \mathbf{g}(\mathbf{a}) + \mathbf{g}(\mathbf{b})$ and $g_k(\mathbf{a}+\mathbf{b}) = g_k(\mathbf{a}) + g_k(\mathbf{b})$, where $\mathbf{b}$ is another vector with size equals to $|S|$.

In the following, denote $\mathbf{e} = [1, 1, ..., 1]'$ as a vector with $|S|$ entries. Also, we choose another $|S|$ vectors with sizes equal to $|S|$, e.g., $\mathbf{e}_1 = [1, 0, ..., 0]'$, $\mathbf{e}_2 = [0, 1, ..., 0]'$ and $\mathbf{e}_{|S|} = [0, 0, ..., 1]'$. Thus, $\mathbf{e} = \mathbf{e}_1 + \mathbf{e}_2 + ... + \mathbf{e}_{|S|}$. In this way, $\mathbf{g}(\mathbf{e}) = \mathbf{g}(\mathbf{e}_1) + \mathbf{g}(\mathbf{e}_2) + ... + \mathbf{g}(\mathbf{e}_{|S|})$ and $g_k(\mathbf{e}) = g_k(\mathbf{e}_1) + g_k(\mathbf{e}_2) + ... + g_k(\mathbf{e}_{|S|})$.

According to the definition, $g_k(\mathbf{e}) = f_{S\to k}$ and $g_k(\mathbf{e}_i) = f_{i\to k}^{S\setminus\{i\}}$. Thus, $f_{S\to k} = \sum_{i\in S} f_{i\to k}^{S\setminus\{i\}}$ holds.

Since $f_{S\to V} = \sum_{k\in V} f_{S\to k}$, we could get

$$f_{S\to V} = \sum_{k\in V}\sum_{i\in S} f_{i\to k}^{S\setminus\{i\}} = \sum_{i\in S}\sum_{k\in V} f_{i\to k}^{S\setminus\{i\}} = \sum_{i\in S} f_{i\to V}^{S\setminus\{i\}}.$$

■

Upper Bounds. Given node set S, we find that the independent influence for each node i ($f_{i\to V}^{S\setminus\{i\}}$, $i \in S$) is no bigger than its original influence ($f_{i\to V}$), and the original influence is no bigger than $p_i = \sum_{k=1}^{n} p_{ki}$.

Theorem 2: $f_{i\to V}^{S\setminus\{i\}} \leq f_{i\to V} \leq p_i$

Proof: First, let us prove $f_{i\to V}^{S\setminus\{i\}} \leq f_{i\to V}$ by $f_{i\to j}^{S\setminus\{i\}} \leq f_{i\to j}$ for $\forall j \in V$. From Definition 1, there is

$$f_{i\to i} = 1$$
$$f_{i\to j} = d\sum_{k\in N_j} t_{kj} f_{i\to k}, \quad for\ j \neq i.$$

[4]Note that the dimension of both $\mathbf{f}_i^{S\setminus\{i\}}$ and $\nu_{\mathbf{i}}^{S\setminus\{i\}}$ can be viewed as $n - |S| + 1$, since for $\forall j \in S and j \neq i$, $f_{i\to j}^{S\setminus\{i\}} = 0$ and $\nu_{\mathbf{i},j}^{S\setminus\{i\}}$=0.

[5]Here is a node set but a single node.

which could be transformed into the following equivalent formation

$$\begin{aligned} f_{i\to i} &= 1, \\ f_{i\to j} &= d\sum_{l\in N_j} t_{lj} f_{i\to l} = a_j, \quad for\ j \in S \,\&\, j \neq i, \\ f_{i\to k} &= d\sum_{l\in N_k} t_{lk} f_{i\to l}, \quad for\ k \notin S, \end{aligned}$$

where a_j must be a number no less than 0 (for d, t_{lj} and $f_{i\to l}$ are all no less than 0). Following the notations in Theorem 1, there is $\mathbf{f}_i = \mathbf{g}(\mathbf{a}_i)$, where $\mathbf{a}_i = [a_1, a_2, ...a_i = 1, ...a_{|S|}]'$[6].

Meanwhile, $\mathbf{f}_i^{S\setminus\{i\}} = \mathbf{g}(\mathbf{e}_i)$. Then, we have

$\mathbf{f}_i = \mathbf{f}_i^{S\setminus\{i\}} + (\mathbf{g}(\mathbf{a}_i) - \mathbf{g}(\mathbf{e}_i)) = \mathbf{f}_i^{S\setminus\{i\}} + \mathbf{g}(\mathbf{a}'_i)$, where $\mathbf{a}'_i = \mathbf{a}_i - \mathbf{e}_i = [a_1, a_2, ...a_i = 0, ...a_{|S|}]' \geq \mathbf{0}$. Since $\mathbf{g}(\mathbf{a}'_i) \geq \mathbf{g}(\mathbf{0}) = \mathbf{0}$, $\mathbf{f}_i \geq \mathbf{f}_i^{S\setminus\{i\}}$.

That is $\forall j \in V$, $f_{i\to j} \geq f_{i\to j}^{S\setminus\{i\}}$ holds.

Second, $f_{i\to V} \leq p_i$ can be proved in the following way. By Equation (4), we have $\mathbf{P}^{-1}\mathbf{f}_i = (\mathbf{I} - d\mathbf{T}')\mathbf{f}_i = \nu_{\mathbf{i}}$. Thus $1 - d\sum_{k\neq i} t_{ki} f_{i\to k} = \nu_{\mathbf{i},i}$.

Since both $t_{ki} \geq 0$ and $f_{i\to k} \geq 0$, we can get $\nu_{\mathbf{i},i} \leq 1$. Meanwhile, as $f_{i\to k} = p_{ki}\nu_{\mathbf{i},i}$, $f_{i\to k} \leq p_{ki}$.

Thus, $f_{i\to V} = \sum_{k=1}^{n} f_{i\to k} \leq p_i$ holds.

In this way, $f_{i\to V}^{S\setminus\{i\}} \leq f_{i\to V} \leq p_i$ holds.

■

B. Applications

Given a set of seed nodes S (e.g., the seeds selected by CELF [6] or PMIA [8] for viral marketing), we could evaluate the proposed independent influence computation by a number of applications. With the help of the above two properties, in this paper, we choose two of the most important and intuitive ones: rank the seeds based on their independent influence, and quickly find the Top-K influential seed nodes from S.

Seeds Ranking. From Theorem 1, we can see that the influence of set S is actually the sum of each node's independent influence, which means the system or the agent could figure out the real influence contribution of each selected seed. Thus, the system can rank and pay the seeds based on their independent influence, or further remove the ones which borrow lots of influence from other seeds in current seed set.

Specifically, these independent influence could be computed by Equation (11) and Equation (12). Let's take the node $i \in S$ as an example. For computing i's independent influence, we first compute the $\nu_{\mathbf{i},i}^{S\setminus\{i\}}$ and $\mathbf{f}_i^{S\setminus\{i\}}$ in Equation (11) by the Gauss-Seidel method, which is similar to Algorithm 1 in $O(|A|)$ except that we will solve $(\mathbf{I} - d\mathbf{T}'_{\overline{S^i S^i}})\Gamma_{\cdot i} = \mathbf{e}_i$ at this time. Then, the independent influence of i, $f_{i\to V}^{S\setminus\{i\}}$, could be summarized by Equation (12). Under the linear definition, the above procedure could be run for all the nodes in S in $O(|S||A|)$. Finally, we can better understand each seed, e.g., by ranking them based on their independent influence.

Top-K Influential Seeds Identification. In some scenarios, the seed set S is usually very large, and at this time we are more interested in the seeds at both ends, i.e., finding the Top-K or Bottom-K independent influential seeds from S. Actually, in this paper, we focus on quickly identifying the Top-K influential seeds, and leave the Bottom-K identification problem for future work.

The most straightforward way to select the Top-K influential seeds is computing the independent influence for each seed, and then ranking them. However, as illustrated before, this will take $O(|S||A|)$ for our linear definition, and much more time consuming for IC model. Luckily, Theorem 2 provides two upper bounds for independent influence, and one of them can be used to develop an efficient algorithm. Specifically, the vector $\boldsymbol{p} = \mathbf{P}'\mathbf{e} = [p_1, ..., p_n]'$ contains the upper bounds (e.g., p_j) for all the seeds[7], and it could be finished in $O(|A|)$ by the Gauss-Seidel method. This computation is also similar to Algorithm 1 except that we will solve

$$(\mathbf{I} - d\mathbf{T})\boldsymbol{p} = \mathbf{e}, \tag{17}$$

as $(\mathbf{P}')^{-1}\boldsymbol{p} = \mathbf{e}$ and $\mathbf{P}$=$(\mathbf{I} - d\mathbf{T}')^{-1}$. It is worth noting that the computation of PageRank values follows the same procedure and time complexity.

Then, these upper bounds are used to save computations, and the entire framework is shown in Algorithm 2[8]. In a nutshell, if we only have to compute the independent influence value for N seeds (i.e., we have to search for N candidate nodes), the time complexity of Algorithm 2 is $O((N+1)|A|)$. From the experiments, we can see that usually $N << |S|$ when K is small, which means our upper bounds are effective.

V. Experimental Results

In this section, we provide empirical validation on several networks. Specifically, we demonstrate: 1) Two case studies, which illustrate that it is important to distinguish the independent influence of each node i by removing other seeds i.e., $S\setminus\{i\}$; 2) The effectiveness and efficiency of our method on ranking the seeds by independent influence computation; 3) The effectiveness of our upper bound for quickly identifying the Top-K influential seeds (i.e., Algorithm 2).

A. Experimental Setup

We conduct experiments on the following five datasets:

- **MovieLens** is a movie consumption network [21] that we constructed from the MovieLens dataset[9];
- **Polblogs** is a network showing the links between politician blogs [22];

[6]Without loss of generality, we set $S = \{1, 2, ...|S|.\}$

[7]Actually, the upper bounds of other nodes can be also computed.

[8]This way of using upper bounds is similar to that in [10].

[9]http://www.grouplens.org/node/73

Algorithm 2: Top-K Independent Influential Seeds Identification

input : $G = (V, A, \mathbf{T})$;
d is the damping factor;
S is the given set of seed nodes;
K is number of the most influential seeds.
output: $\mathbf{S}_K$: Top-K independent influential seeds.
$\mathbf{S}_K$=∅;
Compute $\boldsymbol{p} = [p_1, ..., p_n]'$ in $O(|A|)$; *//Equation (17)*
for *each node i in S* **do**
  $U_i = p_i$; *// Upper bound*
  $IsBound_i = True$;
while $|\boldsymbol{S}_K| < K$ **do**
  Find node i with the biggest U_i in S;
  if $IsBound_i == True$ **then**
    Compute $f_{i\to k}^{S\setminus\{i\}} = \frac{\Gamma_{ki}}{\Gamma_{ii}}$ for all ks in $O(|A|)$;
    //Solve $(\mathbf{I} - d\mathbf{T}'_{\overline{S^i S^i}})\Gamma_{\cdot i} = \mathbf{e}_i$ by Algorithm 1
    $U_i = f_{i\to V}^{S\setminus\{i\}}$; *//Equation (12)*
    $IsBound_i = False$;
  else
    $\mathbf{S}_K = \mathbf{S}_K \bigcup i$;
    U_i =MINIM; *//E.g., 0*
return $\boldsymbol{S}_K$;

Table II
STATISTICS OF FIVE NETWORKS.

Networks	MovieLens	Polblogs	DBLP-DM	Epinion	Amazon
#Nodes	1,682	1,490	53,872	75,879	262,111
#Edges/Arcs	312,400	19,090	160,968	508,837	1,234,877

- **DBLP-DM** is a scientific collaboration network from DBLP[10]. We select the research papers published before Jan. 2013 in several typical top-ranked journals (e.g., DMKD,TKDE) and conferences (e.g., KDD, ICDM, SDM) in data mining, and the authors are used as nodes to construct the scientific collaboration network;
- **Epinion**[11] is a who-trust-whom online social network of a general consumer review site Epinions.com;
- **Amazon**[12] is a co-purchase network of the products from Amazon.com.

Some basic statistics about these directed networks are given in Table II. Note that we use MovieLens and Polblogs mainly for case studies, since they are comparably small and the time-consuming greedy algorithm CELF could be finished quickly. For experiments, the transmission matrix $\mathbf{T}$ is set the same as the transmission matrix of WC model [16], i.e., t_{ij} on edge(j, i) equals to $\frac{Weight(\mathcal{A}_{ji})}{OutWeight(j)}$.

Benchmark Methods. In the following, we call our method as **Linear**, and we compare Linear with several benchmark methods:

- **Degree** measures the independent influence based on the node's degree or its DegreeDiscount value [7]. Each time the best result of these two metrics are chosen for comparison.
- **InfluenceRank** is a method recently proposed in [9], where the independent influence is measured by the InfluenceRank value.
- **PageRank** [20] measures the independent influence of each node by the independent PageRank value.

For computing the independent influence, all the benchmarks are run after removing the given nodes ($S\setminus\{i\}$) from the network. Note that, to the best of our knowledge, none of these benchmarks follow the similar properties that we have found in Linear (i.e., Theorem 1 and Theorem 2).

The following experiments are conducted on the same platform. For the purpose of comparison, we record the best performance of each algorithm by tuning their parameters, e.g., the damping factor d in InfluenceRank, PageRank and Linear is set to be 0.85, and the propagation probability $p = 0.01$ for DegreeDiscount [7].

Evaluation Metrics. Since it is hard for measuring the real influence of each node, we refer to the result output by WC model [16] as the ground truth. The major reason is that as a kind of IC model, WC model is the most widely accepted influence computation model, and it could simulate the real-world influence propagation process more accurately than the Uniform IC Model [17]. Specifically, we run Monte-Carlo simulation under the WC model for sufficiently many (i.e., 20,000) times, and each time we sum up the influence spread (i.e., the expected number of nodes that will be influenced) on the network, then the average influence spread is used for estimating the real influence. Meanwhile, for making more meaningful and persuasive illustrations, in this paper we mainly focus on evaluating the node ranking of each method based on their output independent influence. In other words, the better methods could output the ranking list more similar to that output by WC model.

B. Case Studies

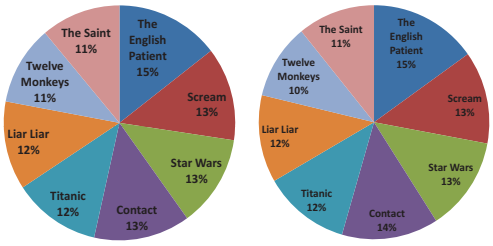
In this section, we use two case studies to illustrate the importance of measuring independent influence. Specifically, the first case study illustrates that if we select a set of seeds S for viral marketing, the contribution (independent influence) of each single seed is quite different. The second case study illustrates that the seed's independent influence is affected by other seeds (i.e., $S\setminus\{i\}$).

Case Study 1. In this case study, we first use CELF method (a greedy algorithm based on IC model) [6] to select a set of seed nodes (i.e., $|S| = 8$) for viral marketing. Then we show the percentage of the independent influence of each seed in this set. Also, we re-rank the seeds based on their independent influence computed by WC model and Linear model, respectively. This experiment is performed on two small datasets MovieLens and Polblogs, since CELF is very

[10]http://dblp.uni-trier.de/xml/
[11]http://snap.stanford.edu/data/soc-Epinions1.html
[12]http://snap.stanford.edu/data/amazon0302.html

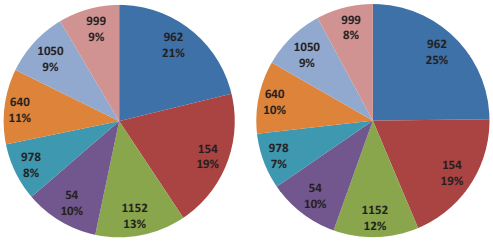
Table III
CASE STUDY OF THE SEED SET (TOP-8) IN MOVIELENS (MOVIE NAME) AND POLBLOGS (NODE ID).

Data	Alg.	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
MovieLens	CELF	The English Patient	Scream	Star Wars	Contact	Titanic	Liar Liar	Twelve Monkeys	The Saint
	WC	The English Patient(*59.1*)	Contact(*54.8*)	Scream(*52.4*)	Titanic(*49.8*)	Star Wars(*51.7*)	Liar Liar(*49.6*)	The Saint(*45.3*)	Twelve Monkeys(*44.6*)
	Linear	The English Patient(*34.9*)	Contact(*31.3*)	Star Wars(*30.1*)	Scream(*29.6*)	Titanic(*28.1*)	Liar Liar(*27.9*)	The Saint(*25.0*)	Twelve Monkeys(*24.1*)
Polblogs	CELF	962	154	1152	54	978	640	1050	999
	WC	962(*87.0*)	154(*79.7*)	1152(*52.0*)	640(*42.9*)	54(*42.6*)	1050(*37.8*)	999(*35.0*)	978(*33.2*)
	Linear	962(*84.1*)	154(*63.9*)	1152(*40.1*)	640(*34.2*)	54(*33.7*)	1050(*30.0*)	999(*26.7*)	978(*26.4*)



(a) WC model (b) Linear model

Figure 2. The independent influence pie graph of Movielens seeds.



(a) WC model (b) Linear model

Figure 3. The independent influence pie graph of Polblogs seeds.

time consuming. The final results are shown in Table III and Figure 2 and Figure 3, respectively.

For each dataset, the first row in Table III are the seeds selected by CELF, and the rank orders of these seeds are also given. Then, the second and the third row rank these seeds based on their independent influence ($f_{i\to V}^{S\setminus\{i\}}$, the number in each bracket (.) of this specific seed) computed by WC model and linear model, respectively. Let's take MovieLens as an example, we can see the seeds selected by CELF are well-known movies and they cover a number of movie genres. Thus, they may lead to the maximization of influence spread. However, the rank of the nodes based on their independent influence (independent from other seven seeds) is quite different from the CELF selection. For instance, movie "Contact" is actually more influential (Rank 2) under both WC model and Linear model. Meanwhile, we could see that the ranking lists of WC model and Linear model are quite similar. For deeper understanding, we quantify each node's independent influence, and normalize them into pie graphs[13](Figure 2 and Figure 3). From these two figures we have the following interesting observations: First, the independent influence varies a lot for different seeds. This implies that the seed set S selected by CELF can not guarantee that each seed is useful, and it is necessary to figure out the seed's real influence; Second, the pie graphs got by WC model and Linear model are quite similar to each other, i.e., the Linear estimation is consistent with the estimation of WC model.

Case study 2. This is an even more straightforward yet complex case study, where the most influential and independent influential nodes in two datasets (MovieLens and DBLP-DM) are illustrated. In this case study, we first use different methods (i.e., WC, Linear, Degree, InfluenceRank, PageRank) select the Top-8 influential nodes. Then, we manually choose two set of nodes (5 movies or researchers as $S\setminus\{i\}$) from each dataset and find the most influential nodes independent from the nodes in these two sets. The final results are illustrated in Table IV and Table V.

In the first five rows (row 1-5) of each table are the most influential nodes (at this time $S\setminus\{i\}=\emptyset$, i.e., $S=\{i\}$) selected by each method, and row 6-10 and row 11-15 are the most independent influential nodes with respect to two seed sets, respectively. Let's take Table V as an example, all the researchers in first five rows are famous researchers. Though the algorithms are quite different from each other, the influential researchers determined are quite similar. Meanwhile, the researchers' independent influence is affected by the given seeds ($S\setminus\{i\}$). It is hard to find useful information in row 6-10 and row 11-15 at the first glance, however, some interesting observations could be explained. For instance, in our collected data, both Dr. Ming-Syan Chen and Dr. Charu C. Aggarwal have close collaboration with Dr. Philip S. Yu, thus when Philip S. Yu is chosen in the given seed set, the independent influence of Ming-Syan Chen and Charu C. Aggarwal is affected a lot (this can be seen from their ranking orders in row 6-10). In contrast, when Ming-Syan Chen and Charu C. Aggarwal are included in the given seed set, Philip S. Yu will lost some independent influence (row 11-15). In summary, this case study illustrates that the nodes' independent influence is affected by other nodes in the seed sets, and the closer the nodes the more influence will be lost. Meanwhile, different algorithms output different node ranking lists, since they compute independent influence following different strategies.

C. Independent Influence Computation

In this section, we present the performance comparison on seeds ranking by independent influence computation between Linear and the benchmark methods. Specifically, we first generate a seed set S by randomly selecting given number ($|S|$=20, 40, 60 or 80) of seeds from the Top-

[13]It is worth noting that the total influence and independent influence output by WC model do not follow Theorem 1.

Table IV
MOST INFLUENTIAL OR INDEPENDENT INFLUENTIAL (GIVEN TWO DIFFERENT SETS) NODES (MOVIES) IN MOVIELENS.

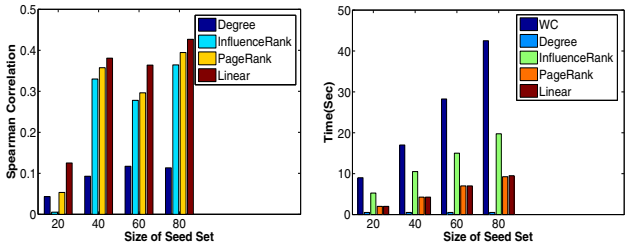
Given ($S \setminus \{i\}$)	Alg.	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
	WC	The English Patient	Contact	Star Wars	Liar Liar	Fargo	Scream	L.A. Confidential	Air Force One
	Linear	The English Patient	Contact	Star Wars	Scream	Fargo	Titanic	Liar Liar	Air Force One
	Degree	Star Wars	Contact	Fargo	Return of the Jedi	Liar Liar	Scream	The English Patient	Toy Story
	InfluenceRank	The English Patient	Contact	Star Wars	Fargo	Scream	Titanic	Liar Liar	Return of the Jedi
	PageRank	The English Patient	Contact	Star Wars	Scream	Fargo	Titanic	Liar Liar	Air Force One
{Star Wars	WC	Scream	Titanic	Chasing Amy	L.A. Confidential	Air Force One	The Fully Monty	Return of the Jedi	Conspiracy Theory
Liar Liar	Linear	Scream	Titanic	Air Force One	The Fully Monty	L.A. Confidential	Return of the Jedi	Chasing Amy	Evita
Contact	Degree	Return of the Jedi	Scream	Toy Story	Twelve Monkeys	Air Force One	The Godfather	Independent Day(ID4)	Pulp Fiction
Fargo	InfluenceRank	Scream	Titanic	Return of the Jedi	The Full Monty	L.A. Confidential	Air Force One	Twelve Monkeys	The Godfather
The English Patient}	PageRank	Scream	Titanic	Air Force One	The Full Monty	L.A. Confidential	Return of the Jedi	Chasing Amy	Evita
{Scream	WC	The English Patient	Contact	Star Wars	Liar Liar	Fargo	The Fully Monty	L.A. Confidential	The Game
Air Force One	Linear	The English Patient	Contact	Star Wars	Fargo	Liar Liar	The Fully Monty	L.A. Confidential	Chasing Amy
Titanic	Degree	Star Wars	Contact	Fargo	Liar Liar	The English patient	Twelve Monkeys	The Godfather	Independence Day(ID4)
Toy Story	InfluenceRank	The English Patient	Contact	Fargo	Star Wars	Liar Liar	The Fully Monty	L.A. Confidential	Twelve Monkeys
Return of the Jedi}	PageRank	The English Patient	Contact	Star Wars	Fargo	Liar Liar	The Fully Monty	L.A. Confidential	Chasing Amy

Table V
MOST INFLUENTIAL OR INDEPENDENT INFLUENTIAL (GIVEN TWO DIFFERENT SETS) NODES (RESEARCHERS) IN DBLP-DM.

Given ($S \setminus \{i\}$)	Alg.	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
	WC	P. S. Yu	J. Han	C. Faloutsos	H. Wang	E. J. Keogh	C. C. Aggarwal	J. Pei	K. Wang
	Linear	P. S. Yu	J. Han	C. Faloutsos	M. Chen	Eamonn J. Keogh	K. Wang	J. Pei	C. C. Aggarwal
	Degree	J. Han	P. S. Yu	C. Faloutsos	J. Pei	Q. Yang	E. J. Keogh	K. Wang	H. Mannila
	InfluenceRank	P. S. Yu	J. Han	C. Faloutsos	M. Chen	E. J. Keogh	K. Wang	J. Pei	C. C. Aggarwal
	PageRank	P. S. Yu	J. Han	C. Faloutsos	M. Chen	E. J. Keogh	K. Wang	J. Pei	C. C. Aggarwal
	WC	H. Mannila	H. Kargupta	V. Kumar	J. Pei	P. Melville	T. Jiang	N. Abe	J. Sun
{P. S. Yu, J. Han	Linear	H. Mannila	Q. Yang	K. Wang	E. Bertino	S. Parthasarathy	V. Kumar	H. Kriegel	M. Chen
C. Faloutsos	Degree	Q. Yang	H. Mannila	Z. Chen	H. Xiong	H. Kriegel	M. J. Zaki	J. X. Yu	V. Kumar
E. J. Keogh, H. Wang}	InfluenceRank	M. Chen	K. Wang	Q. Yang	S. Parthasarathy	E. Bertino	H. Mannila	H. Kargupta	J. Pei
	PageRank	H. Mannila	Q. Yang	K. Wang	E. Bertino	S. Parthasarathy	V. Kumar	H. Kriegel	M. Chen
	WC	J. Han	C. Faloutsos	P. S. Yu	A. Gionis	E. J. Keogh	T. Jiang	H. Mannila	J. Wang
{M. Chen, J. Pei	Linear	J. Han	P. S. Yu	C. Faloutsos	E. J. Keogh	H. Mannila	E. Bertino	T. Li	P. Smyth
C. C. Aggarwal	Degree	C. Faloutsos	J. Han	E. J. Keogh	H. Mannila	H. Xiong	W. Fan	J. X. Yu	H. Kriegel
K. Wang, Q. Yang}	InfluenceRank	J. Han	P. S. Yu	C. Faloutsos	E. J. Keogh	E. Bertino	H. Mannila	S. Parthasarathy	P. Smyth
	PageRank	J. Han	P. S. Yu	C. Faloutsos	E. J. Keogh	H. Mannila	E. Bertino	T. Li	P. Smyth

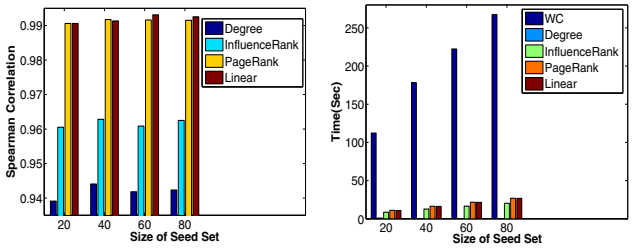
100 nodes with highest degree. Then, we rank these seeds by Linear and the benchmarks, respectively. Finally, we compute and compare the Spearman correlations (the bigger the better)[14] between these rankings with the ground truth (i.e., 20,000 times Monte-Carlo simulation of WC model). Meanwhile, we record the running time. With respect to each dataset, the above procedure will be run 4 times for each size of seed set, and the average results are used for final comparison.

The experimental results on DBLP-DM, Epinion and Amazion are shown in Figure 4, Figure 5 and Figure 6, respectively. We could find the similar observations from these figures: For the effectiveness comparison, the rankings of Linear method have the highest Spearman Correlation values with the ground truth for most of the time, while Degree based methods perform the worst since the nodes with the highest degree may not be most influential and vice versa; For the efficiency issue, we could see that it is most time consuming to run the WC model and our Linear method is as fast as PageRank, while degree based methods are very efficient since they only need to search the nodes once. In summary, Linear method could effectively and efficiently rank the seeds based on their independent influence.



(a) Ranking result comparison (b) Running time comparison

Figure 4. Independent influence computation on DBLP-DM.



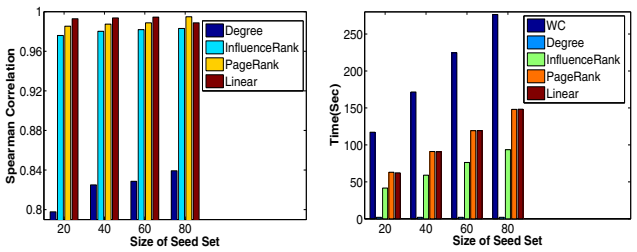
(a) Ranking result comparison (b) Running time comparison

Figure 5. Independent influence computation on Epinion.

D. Top-K Influential Seeds Identification

In this section, we provide empirical validation on the second application, Top-K independent influential seeds identification from seed set S. The experimental setup is similar to that in Section V-C: First, we generate a seed set S by randomly selecting given number ($|S|$=20, 40, 60 or 80) of seeds from the Top-100 nodes with highest degree. Then, we identify the Top-10 (i.e., we fix K=10)

[14]http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient



(a) Ranking result comparison (b) Running time comparison

Figure 6. Independent influence computation on Amazon.

independent influential seeds from each of these sets by Linear (Algorithm 2) and the benchmarks. For comparing the effectiveness of each method, we compute the Jaccard similarity (the bigger the better)[15] of their outputs with the ground truth (i.e., Top-10 seeds output by the 20,000 times Monte-Carlo simulation of WC model). Also, this procedure is run 4 times for each size of seed set on each dataset, and the average values are used for final comparison.

Besides computing the Jaccard similarity of the Top-10 seeds, we also compare each method's running time and the effectiveness of our proposed upper bounds in Algorithm 2. Along this line, the experimental results on DBLP-DM, Epinion and Amazion can be found in Figure 7, Figure 8 and Figure 9, respectively. For the Jaccard similarity comparison (the leftmost subfigures), we can see that the most influential seeds identified by Linear have the largest overlaps with the ground truth, with the average Jaccard similarity value bigger than 0.79. For the running time comparison (the middle subfigures), WC model and InfluenceRank are the most and second most time consuming algorithms. Among these algorithms, Degree is the fastest algorithm, however, it performs the worst with respect to Top-10 seeds identification (Figure 7(a), Figure 8(a) and Figure 9(a)). Meanwhile, in this experiment Linear becomes much faster than PageRank, and this is due to the help of our upper bounds. More specifically, we demonstrate the effectiveness of our upper bounds by presenting the number of searched nodes (i.e., N) for finding the Top-10 independent influential seeds in each seed set (the rightmost subfigures). We can observe that this number is comparably small with respect to the size of the entire seed set, which also indicates that Algorithm 2 is scalable.

VI. Concluding Remarks And Future Work

In this paper, we provided a focused study on measuring the node's independent influence by a linear social influence model. Along this line, we first presented the definition of linear social influence model and the independent influence in detail. Then, we found two properties of the proposed independent influence, i.e., the influence of a set of nodes (this node set may be selected manually or by some algorithms, e.g., CELF) is actually the sum of their independent influence and the independent influence is no bigger than each seed's original influence. Moreover, we applied this independent influence computation for seeds ranking and quickly identifying Top-K independent influential seeds from the seed set. Finally, an empirical study was conducted on five network datasets, and the results demonstrated the effectiveness and efficiency of the proposed independent influence measures.

This paper provides an introduction of the problem space in independent social influence analysis. The area is still in its infancy, and we anticipate that more techniques will be developed. Specifically, in the future, we plan to find more reasonable metrics on influence evaluation. Meanwhile, quickly identifying the Bottom-K influential seeds and the topic-sensitive [23] independent social influence computations are also possible directions for future research.

[15]http://en.wikipedia.org/wiki/Jaccard_index

VII. Acknowledgements

The work described in this paper was supported by grants from Natural Science Foundation of China (Grant No. 61073110), Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20113402110024), and National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2012BAH17B03). Enhong Chen gratefully acknowledges the support of Huawei Technologies Co., Ltd. (Grant No. YBCB2012086).

References

[1] C. C. Aggarwal, *An introduction to social network data analytics*. Springer, 2011.

[2] J. Sun and J. Tang, "A survey of models and algorithms for social influence analysis," in *Social Network Data Analytics*. Springer, 2011, pp. 177–214.

[3] P. Domingos and M. Richardson, "Mining the network value of customers," in *SIGKDD*. ACM, 2001, pp. 57–66.

[4] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[5] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, pp. 1420–1443, 1978.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *SIGKDD*. ACM, 2007, pp. 420–429.

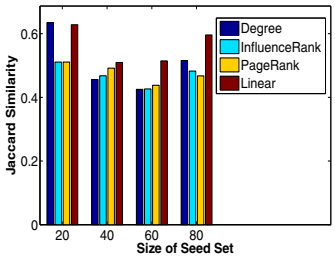
[7] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *SIGKDD*. ACM, 2009, pp. 199–208.

[8] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *SIGKDD*. ACM, 2010, pp. 1029–1038.

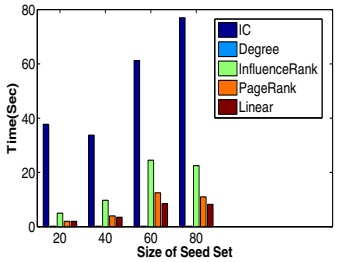
[9] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks," in *ICDM*. IEEE, 2012, pp. 918–923.

[10] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective," in *IJCAI*, 2013, pp. 2740–2746.

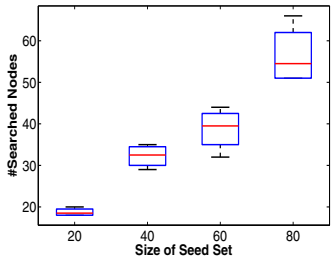
[11] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *SIGKDD*. ACM, 2008, pp. 7–15.



(a) Jaccard similarity of Top-10 seeds

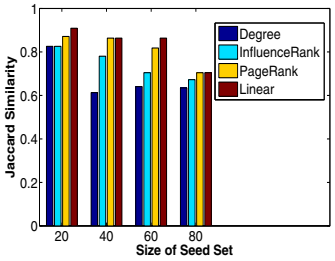


(b) Running time comparison

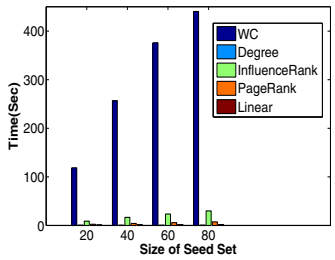


(c) Effectiveness of the upper bounds

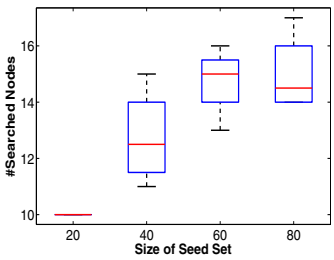
Figure 7. Top-10 independent influential seeds identification results on DBLP.



(a) Jaccard similarity of Top-10 seeds

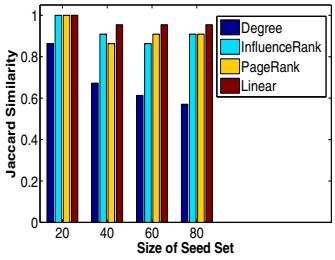


(b) Running time comparison

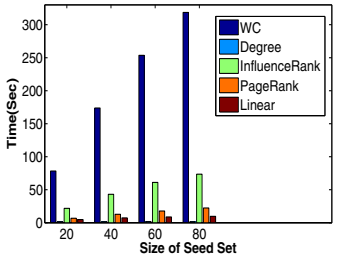


(c) Effectiveness of the upper bounds

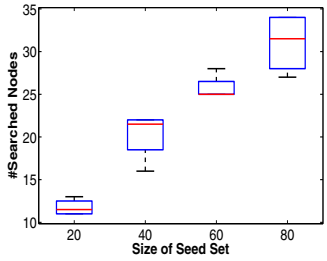
Figure 8. Top-10 independent influential seeds identification results on Epinion.



(a) Jaccard similarity of Top-10 seeds



(b) Running time comparison



(c) Effectiveness of the upper bounds

Figure 9. Top-10 independent influential seeds identification results on Amazon.

[12] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *WSDM*. ACM, 2010, pp. 241–250.

[13] G. Ver Steeg and A. Galstyan, "Information-theoretic measures of influence based on content dynamics," in *WSDM*. ACM, 2013, pp. 3–12.

[14] C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks," in *SDM*, 2011, pp. 522–533.

[15] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *PKDD*. Springer, 2006, pp. 259–271.

[16] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD*. ACM, 2003, pp. 137–146.

[17] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. A. Shad, "On approximation of real-world influence spread," in *PKDD*. Springer, 2012, pp. 548–564.

[18] P. Li, J. X. Yu, H. Liu, J. He, and X. Du, "Ranking individuals and groups by influence propagation," in *PAKDD*. Springer, 2011, pp. 407–419.

[19] D.-N. Yang, H.-J. Hung, W.-C. Lee, and W. Chen, "Maximizing acceptance probability for active friending in on-line social networks," *arXiv preprint arXiv:1302.7025*, 2013.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[21] Q. Liu, B. Xiang, E. Chen, Y. Ge, H. Xiong, T. Bao, and Y. Zheng, "Influential seed items recommendation," in *RecSys*. ACM, 2012, pp. 245–248.

[22] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[23] T. H. Haveliwala, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *IEEE TKDE*, vol. 15, no. 4, pp. 784–796, 2003.

Appendix 1

In this section, we prove that $\mathbf{g}(\mathbf{a}+\mathbf{b}) = \mathbf{g}(\mathbf{a}) + \mathbf{g}(\mathbf{b})$ and $g_k(\mathbf{a}+\mathbf{b}) = g_k(\mathbf{a}) + g_k(\mathbf{b})$ for Theorem 1.

First, we rewrite $\mathbf{g}(\mathbf{a})$ as

$g_i(\mathbf{a}) = d\sum_{k=1} t_{ki} g_k(\mathbf{a}) + \nu_i(\mathbf{a})$ for $i = 1, 2..., n$.

When $i \notin S$, $\nu_i(\mathbf{a})$ equals to 0. Otherwise, $\nu_i(\mathbf{a})$ is a value to make sure $g_i(\mathbf{a}) = a_i$.

Similar to Equation (3) and Equation (13), we have $\mathbf{g}(\mathbf{a}) = (\mathbf{I} - d\mathbf{T}')^{-1}\nu(\mathbf{a}) = \mathbf{P}\nu(\mathbf{a})$, and $\nu(\mathbf{a}) = \mathbf{P}^{-1}\mathbf{g}(\mathbf{a})$.

Then, just considering the nodes in S, we could get $\nu_S(\mathbf{a}) = \mathbf{P}_{SS}^{-1}\mathbf{a}$, where $\nu_S(\mathbf{a})$ and $\mathbf{P}_{SS}^{-1}$ are the remaining entries in $\nu(\mathbf{a})$ and $\mathbf{P}^{-1}$, respectively.

Since $\nu_S(\mathbf{a})+\nu_S(\mathbf{b})=\mathbf{P}_{SS}^{-1}\mathbf{a} + \mathbf{P}_{SS}^{-1}\mathbf{b}=\mathbf{P}_{SS}^{-1}(\mathbf{a}+\mathbf{b})=\nu_S(\mathbf{a}+\mathbf{b})$, thus, $\mathbf{g}(\mathbf{a}+\mathbf{b}) = \mathbf{P}\nu(\mathbf{a}+\mathbf{b}) = \mathbf{g}(\mathbf{a}) + \mathbf{g}(\mathbf{b})$ holds, and incorrespondingly, $g_k(\mathbf{a}+\mathbf{b}) = g_k(\mathbf{a}) + g_k(\mathbf{b})$.