2012 IEEE 12th International Conference on Data Mining

IRIE: Scalable and Robust Influence Maximization in Social Networks

Kyomin Jung
KAIST
Republic of Korea
kyomin@kaist.edu

Wooram Heo
KAIST
Republic of Korea
modesty83@kaist.ac.kr

Wei Chen
Microsoft Research Asia
China
weic@microsoft.com

***Abstract*—Influence maximization is the problem of selecting top k seed nodes in a social network to maximize their influence coverage under certain influence diffusion models. In this paper, we propose a novel algorithm IRIE that integrates the advantages of influence ranking (IR) and influence estimation (IE) methods for influence maximization in both the independent cascade (IC) model and its extension IC-N that incorporates negative opinion propagations. Through extensive experiments, we demonstrate that IRIE matches the influence coverage of other algorithms while scales much better than all other algorithms. Moreover IRIE is much more robust and stable than other algorithms both in running time and memory usage for various density of networks and cascade size. It runs up to two orders of magnitude faster than other state-of-the-art algorithms such as PMIA for large networks with tens of millions of nodes and edges, while using only a fraction of memory.**

***Keywords*-social network mining, social network analysis, influence maximization, independent cascade model, viral marketing**

I. Introduction

Word-of-mouth or viral marketing has long been acknowledged as an effective marketing strategy. The increasing popularity of online social networks such as Facebook and Twitter provides opportunities for conducting large-scale online viral marketing in these social networks. Two key technology components that would enable such large-scale online viral marketing are modelling influence diffusion and influence maximization. In this paper, we focus on the second component, which is the problem of finding a small set of k *seed nodes* in a social network to maximize their *influence spread* — the expected total number of activated nodes after the seed nodes are activated, under certain influence diffusion models.

In particular, we study influence maximization under the popular independent cascade (IC) model [1] and its extension IC-N model incorporating negative opinions [2]. IC model is one of the most common information diffusion model which is widely used in economics, epidemiology, sociology, and so on [1], [3]. Most of existing researches for the influence maximization problem are based on the IC model, assuming dynamics of information diffusion among individuals are independent. Kempe et al. originally proposed the IC model and a greedy approximation algorithm to solve the influence maximization problem under the IC model [1]. The greedy algorithm proceeds in rounds, and in each round one node with the largest marginal contribution to influence spread is added to the seed set. However, computing influence spread given a seed set is shown to be #P-hard [4], and thus the greedy algorithm has to use Monte-Carlo simulations with a large number of simulation runs to obtain an accurate estimate of influence spread, making it very slow and not scalable. A number of follow-up works tackle the problem by designing more efficient and scalable optimizations and heuristics [3], [5], [6], [7], [4], [6], [8], [9]. Among them PMIA [4] algorithm has stood out as the most efficient heuristic so far, which runs three orders of magnitude faster than the optimized greedy algorithm of [5], [7], while maintaining good influence spread in par with the greedy algorithm.

In this paper, we propose a novel scalable influence maximization algorithm IRIE, and demonstrate through extensive simulations that IRIE scales even better than PMIA, with up to two orders of magnitude speedup and significant savings in memory usage, while maintaining the same level or even better influence spread than PMIA. We also demonstrate that while the running time of PMIA is very sensitive to structural properties of the network such as the clustering coefficient and the edge density, and to the cascade size, IRIE is much more stable and robust over them and always shows very fast running time. In the greedy algorithm as well as in PMIA, each round a new seed with the largest marginal influence spread is selected. To select this seed, the greedy algorithm uses Monte-Carlo simulations while PMIA uses more efficient local tree based heuristics to estimate marginal influence spread of every possible candidate. This is especially slow for the first round where the influence spread of every node needs to be estimated. Therefore, instead of estimating influence spread for each node at each round, we propose a novel global influence ranking method IR derived from a belief propagation approach, which uses a small number of iterations to generate a global influence ranking of the nodes and then select the highest ranked node as the seed. However, the influence ranking is only good for selecting one seed. If we use the ranking to directly select k top ranked nodes as k seeds, their influence spread may overlap with one another and not result in the best

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.79

overall influence spread. To overcome this shortcoming, we integrate IR with a simple influence estimation (IE) method, such that after one seed is selected, we estimate additional influence impact of this seed to each node in the network, which is much faster than estimating marginal influence for many seed candidates, and then use the results to adjust next round computation of influence ranking. When combining IR and IE together, we obtain our fast IRIE algorithm. Besides being fast, IRIE has another important advantage, which is its memory efficiency. For example, PMIA needs to store data structures related to the local influence region of every node, and thus incurs a high memory overhead. In constrast, IRIE mainly uses global iterative computations without storing extra data structures, and thus the memory overhead is small.

We conduct extensive experiments using synthetic networks as well as five real-world networks with size ranging from $29K$ to $69M$ edges, and different IC model parameter settings. We compare IRIE with other state-of-the-art algorithms including the optimized greedy algorithm, PMIA, simulated annealing (SA) algorithm proposed in [8], and some baseline algorithms including the PageRank. Our results show that (a) for influence spread, IRIE matches the greedy algorithm and PMIA while being significantly better than SA and PageRank in a number of tests; and (b) for scalability, IRIE is some orders of magnitude faster than the greedy algorithm and PMIA and is comparable or faster than SA; and (c) for stability IRIE is much more stable and robust over structural properties of the network and the cascade size than PMIA and the greedy algorithm.

Moreover, to show the wide applicability of our IRIE approach, we also adapt IRIE to the IC-N model, which considers negative opinions emerging and propagating in networks [2]. Our simulation results again show that IRIE has comparable influence coverage while scales much better than the MIA-N heuristic proposed in [2]. The detailed descriptions of IC-N model and our simulation results can be found in the full version of our paper [10].

Related Work. Domingo and Richardson [11] are the first to study influence maximization problem in probabilistic settings. Kempe et al. [1] formulate the problem of finding a subset of influential nodes as a combinatorial optimization problem and show that influence maximization problem is NP-hard. They propose a greedy algorithm which guarantees $(1 - 1/e)$ approximation ratio. However, their algorithm is very slow in practice and not scalable with the network size. In [5], [6], authors propose lazy-foward optimization that significantly speeds up the greedy algorithm, but it still cannot scale to large networks with hundreds of thousands of nodes and edges. A number of heuristic algorithms are also proposed [3], [7], [4], [12], [8] for the independent cascade model. SPM/SP1M of [3] is based on shortest-path computation, and SPIN of [12] is based on Shapley value computation. Both SPM/SP1M and SPIN have been shown to be not scalable [4], [13]. Simulated anneal approach is proposed in [8], which provides reasonable influence coverage and running time. The best heuristic algorithm so far is believed to be the PMIA algorithm proposed by Chen et al. [4], which provides matching influence spread while running at three orders of magnitude faster than the optimized greedy algorithm. PageRank [14] is a popular ranking algorithm for ranking web pages and other networked entities, which considers diffusion processes whose corresponding transition matrix must have column sums equal to one. Hence it can not be directly used for the influence spread estimation. Our algorithm IR overcomes this shortcoming, and uses equations more directly designed for the IC model. More importantly, our IRIE algorithm integrates influence ranking with influence estimation together with the greedy approach, overcoming the general issue of ignoring overlapping influence coverages suffered by all pure ranking methods. Recently, Goyal et al. propose a data-based approach to social influence maximization [15]. Authors defines a new propagation probability model called *call distribution model* that reveals how influence flows in the networks based on datasets and propose a novel algorithm for influence maximization for that model.

II. SETUP

Influence Maximization problem [1] is a discrete optimization problem in a social network that chooses an optimal initial seed set of given size to maximize influence under a certain information diffusion model. In this paper, we consider Independent Cascade (IC) model as the information diffusion process. We first introduce IC model, then provide a formal definition of Influence Maximization problem under the IC model. Let $G = (V, E)$ be a directed graph for a social network and $P_{uv} \in [0, 1]$ be an edge propagation probability assigned to each edge $(u, v) \in E$. Each node represents a user and each edge corresponds to a social relationship between a pair of users. In the IC model, each node has either an active or inactive state and is allowed to change its state from inactive to active, but not the reverse direction.

Given a seed set S, the process of IC model is as follows : At step $t = 0$, all seed nodes $u \in S$ are activated and added to S_0. At each step $t > 0$, a node $u \in S_{t-1}$ tries to affect its inactive out-neighbors $v \in N^{out}(u)$ with probability P_{uv} and all the nodes activated at this step are added to S_t. This process ends at a step t if $|S_t| = 0$. Note that every activated node u belongs to just one of S_i, where $i = 0, 1, ..., t$. Hence, it has a single chance to activate its neighbors $v \in N^{out}(u)$ at the next step that it is activated. This activation of nodes models the spread of information among people by the word-of-mouth effect as a result of marketing campaigns. Under the IC model, let us define our influence function $\sigma(S)$ as the expected number of activated nodes given a seed set.

Formally, *Influence Maximization problem* is defined as follows : Given a directed social network $G = (V, E)$ and P_{uv} for each edge $(u, v) \in E$ and $K \in \mathbb{N}$, the influence maximization problem is to select a seed set $S \subseteq V$ with $|S| = K$ that maximizes influence $\sigma(S)$ under the IC model. In [1], it is shown that the exact computation of optimum solution for this problem is NP-hard, but the Greedy algorithm achieves $(1 - 1/e)$ - approximation by proving the facts that the influence function σ is non-negative, monotone, and submodular. A set function f is called monotone if $f(S) \leq f(T)$ for all $S \subseteq T$, and the definition of submodular function is given in Definition 1.

Definition 1. *A set function $f : 2^V \rightarrow \mathbb{R}$ is submodular if for every $S \subseteq T \subseteq V$ and $v \in V$, $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$.*

Theorem 1 ([1])**.** *For a non-negative, monotone, and submodular influence function σ, let S be a size-K set obtained by the greedy hill-climbing algorithm in Algorithm 1. Then S satisfies $\sigma(S) \geq (1-1/e) \cdot \sigma(S^*)$ where S^* is an optimum solution.*

At each step, Algorithm 1 computes the marginal influence of every node $w \in V \setminus S$ and then add the maximum one into the seed set S until $|S| = K$. Although the greedy algorithm guarantees constant-approximation solutions and is easy to implement, computing the influence function $\sigma(S)$ is proven to be #P-hard [4]. To estimate influence function $\sigma(S)$ efficiently, Monte-Carlo simulation and other heuristics have been used in various previous works [1], [5], [7], [6], [3], [12], [8], [7], [16]. In this paper, we design a novel efficient heuristic algorithm IRIE that estimates the marginal influences of every nodes accurately.

Algorithm 1 Greedy(K)

```
initialize S = ∅
for i ← 1 to K do
    select u ← argmax_{w∈V\S}(σ(S ∪ {w}) − σ(S))
    S = S ∪ {u}
end for
output S
```

III. Our Algorithm

For a given seed set S, let $\sigma(u|S) = \sigma(S \cup \{u\}) - \sigma(S)$. At each round of IRIE, it selects a node u with the largest marginal influence estimate $\sigma(u|S)$. The novelty of our algorithm lies in that we derive a system of linear equations for $\{\sigma(u|S)\}_{u \in V}$ whose solution can be computed fast by an iterative method. Then we use these computed values as our estimates of $\{\sigma(u|S)\}_{u \in V}$.

Simple Influence Rank. We first explain our formula for $\{\sigma(u|S)\}_{u \in V}$ when $S = \emptyset$. Let $\sigma(u) = \sigma(u|\emptyset)$. The basic idea of our formula lies in that the influence of a node u is essentially determined by the influences of u's neighbors under the IC model. First suppose that graph $G = (V, E)$ is a tree graph (we allow tree edges to be bidirectional). For $(v, u) \in E$, we define $m(u, v)$ to be the expected number of activated nodes when $S = \{u\}$ and (u, v) is removed from E. Note that for a tree graph G, $m(u, v)$ is the expected influence from u excluding the direction toward v. Let $\tilde{\sigma}(u)$ and $\tilde{m}(u, v)$ be our estimates of $\sigma(u)$ and $m(u, v)$ respectively. Then we will compute $\tilde{\sigma}(u)$ and $\tilde{m}(u, v)$ from the following formulas.

$$\tilde{\sigma}(u) = 1 + \sum_{v \in N^{out}(u)} P_{uv} \cdot \tilde{m}(v, u), \tag{1}$$

$$\tilde{m}(u, v) = 1 + \left(\sum_{w \in N^{out}(u), w \neq v} P_{uw} \cdot \tilde{m}(w, u) \right). \tag{2}$$

Note that equation (2) forms a system of $|E|$ linear equations on $|E|$ variables. When G is a tree, (2) has a unique solution. We prove the following theorem, whose proof is in [10].

Theorem 2. *For any tree graph, for each node u, $\tilde{\sigma}(u) = \sigma(u)$, and for each edge $(v, u) \in E$, $\tilde{m}(u, v) = m(v, u)$.*

Even when G is not a tree, we can define the same equations (1) and (2). In this case, the $\tilde{\sigma}(u)$ computed from (1) and (2) corresponds to the influence of u when we allow multiple counts of influence from u to each node via different paths. Note that this approach has a similarity with the popular *Belief Propagation (BP)* algorithm. As in the BP, one natural way to compute the solution of (1) and (2) is using an iterative message passing algorithm. Although this method computes good estimates of $\sigma(u)$ for tree and general graphs, its running time may be slow since one iteration takes $O(\sum_{v \in V} d_{in}(v) \cdot d_{out}(v))$ time where $d_{in}(v)$ and $d_{out}(v)$ is the in-degree and out-degree of v respectively. Note that for a fixed $u \in V$, $m(u, v)$'s are very similar for any $v \in N^{in}(u)$ since $m(u, v)$ only excludes the one direction toward v. We also observe in the numerical simulations on real world networks and synthetic networks, for all $u \in V$, $m(u, v)$'s are almost the same for any $v \in N^{in}(u)$. Hence we substituting one variable $r(u)$ for all the $m(u, v)$, $v \in N^{in}(u)$. Then we obtain our formulas for the simplified expected influence $r(u)$ for $S = \{u\}$ as follows :

$$r(u) = 1 + \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right). \tag{3}$$

Note that equation (3) forms a system of $|V|$ linear equations on $|V|$ variables. Let $X = (r(u))_{u \in V}$, and the influence matrix $A \in \mathbb{R}^{|V| \times |V|}$ be $A_{uv} = P_{uv}$. Let

$B = (1, 1, \ldots, 1)^T \in \mathbb{R}^{|V|}$. Then (3) becomes

$$X = AX + B. \tag{4}$$

If $\lim_{t\to\infty} A^t = 0$, the solution of (4) becomes

$$(I - A)X = B.$$
$$(I + A + A^2 + \cdots)(I - A)X = (I + A + A^2 + \cdots)B.$$
$$\therefore X = B + AB + A^2B + \cdots \tag{5}$$

Note that $(A^t)_{uv}$ is the summation of the expectation of influence paths so that the diffusion process begins from a single node set $\{u\}$ and it activates a node v after exactly t number of iterations when we allow loops in the paths. Hence $(A^t \cdot B)_u$ is equal to the expectation of *relaxed* influence of node u after exactly t number of iterations where *relaxed* means that we allow multiple counts of influence on some nodes and loops in the paths.

Hence, from (5), X_u is the expectation of *relaxed* influence of node i. Note that X_u is an upper bound of $\sigma(u)$ for all $u \in V$. Since we should not allow loops in the influence paths or multi-counts for the computation of $\sigma(u)$, we introduce a damping factor $\alpha \in (0, 1)$ as follows.

$$r(u) = 1 + \alpha \cdot \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right). \tag{6}$$

Note that (6) is equivalent to $X = \alpha AX + B$,
and when $\lim_{t\to\infty} (\alpha A)^t = 0$, the solution of (6) becomes

$$X = B + \alpha AB + \alpha^2 A^2 B + \alpha^3 A^3 B + \cdots. \tag{7}$$

For any $A \in \mathbb{R}^{|V|\times|V|}$, when α is smaller than the inverse of the largest eigenvalue of A, $\lim_{t\to\infty} (\alpha A)^t = 0$. Note that if there is no large spreading in the given IC model, then for all $\alpha \in (0, 1)$, $\lim_{t\to\infty} (\alpha A)^t = 0$. Hence in those cases (7) becomes the solution of (6).

To compute X, we use an iterative computation obtained from (6) as follows. Let $r^{(0)}(u) = 1$ for all $u \in V$, and $r^{(t)}(u) = 1+\alpha\cdot\left(\sum_{v\in N^{out}(u)} P_{uv} \cdot r^{(t-1)}(v)\right)$ for all $u \in V$ and $t = 1, 2, \ldots$. Then we have

$$(r^{(t)}(u))_{u\in V} = B + \alpha AB + (\alpha A)^2 B + \cdots + (\alpha A)^t B.$$

Hence $(r^{(t)}(u))_{u\in V}$ converges exponentially fast to the solution of (6) if $\lim_{t\to\infty} (\alpha A)^t = 0$. Even when there is a large spreading, $(r^{(t)}(u))_{u\in V}$, computes good estimates of $(\sigma(u))_{u\in V}$. We call this iterative computation of $(r^{(t)}(u))_{u\in V}$ the simple Influence Ranking (Simple IR). The running time of simple IR becomes very fast since one iteration of simple IR takes $O(\sum_{v\in V} d_{out}(v))$ time.

One possible approach for influence maximization using simple IR would be selecting top-K seed nodes with the highest $r(u)$. However, simple IR can only compute the influence for individual nodes, and $\sigma(S) \neq \sum_{u\in S} \sigma(u)$ in general due to influence dependency among seed nodes. Hence we propose IRIE as an extension of simple IR to overcome this shortcoming.

Influence Rank Influence Estimation. Now we describe IRIE, which performs an estimation of $\{\sigma(u|S)\}_{u\in V}$ for any given seed set S. Let S be fixed and $AP_S(u)$ be the probability that node u becomes activated after the diffusion process, when the seed set is S. Suppose that we can estimate $AP_S(u)$ by some algorithms. Many known algorithms including MIA and its extension PMIA, and Monte-Carlo simulation can be used for this estimation. We call this part of our algorithm as *Influence Estimation (IE)*.

Then we have the following extension of (6) so that $\{r(u)\}_{u\in V}$ estimates $\{\sigma(u|S)\}_{u\in V}$.

$$r(u) = (1 - AP_S(u)) \cdot \left(1 + \alpha \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right) \right). \tag{8}$$

Note that given $\{AP_S(u)\}_{u\in V}$, (8) is a system of linear equations and is exactly same with (6) when $S = \emptyset$. The factor $(1 - AP_S(u))$ indicates the probability that a node u is not activated by a seed set S and the remaining terms are the same as (6).

Let $D \in \mathbb{R}^{|V|\times|V|}$ be a diagonal matrix so that $D_{uu} = (1 - AP_S(u))$. Then for $X = (r(u))_{u\in V}$, (8) becomes $X = \alpha DAX + DB$. IRIE compute the solution of (8) by an iterative computation as in the simple IR. As in the simple IR, when $\lim_{t\to\infty} (\alpha DA)^t = 0$, the iterative computation of $r(u)$ converges to the solution of (6) exponentially fast. Hence repeating the iterative computation of (8) for constantly many times computes $\{r(u)\}$ which is a good estimate of $\{\sigma(u|S)\}_{u\in V}$. Our stopping criteria for IRIE, i.e., choosing the number of iterations t, is described in section IV. Regarding the choice of α, we found by extensive experiments that the accuracy of IRIE is quite similar for broad range of $\alpha \in [0.3, 0.9]$ for most cases. We suggest a fixed $\alpha = 0.7$ since the IRIE shows almost highest accuracy when $\alpha = 0.7$ for all cases of our experiments.

Now we explain how we estimate $AP_S(u)$. Given a seed set S, we compute the Maximum Influence Out-Aborescence (MIOA) [4] of s for all $s \in S$. By generating MIOA structure for all the seed node $s \in S$, we estimate $AP_S(u)$ according to $AP_S(u) = \sum_{s\in S} AP_s(u)$. Note that the IE part can be replaced with any other algorithm that estimates $AP_S(u)$, making our IRIE algorithm to be a general framework.

IV. Experiments

A. Experimental Setup

1) Datasets: We perform experiments on five real-world social networks, whose edge sizes range from 29K to 69M. First, we have two (undirected) co-authorship network, collected from ArXiv General Relativity ($|V| = 5K, |E| = 29K$) and DBLP Computer Science Biblography Database

Figure IV.2: Running time of algorithms

($|V| = 655K, |E| = 2M$) [17], denoted by ArXiv and DBLP respectively. We also have three (directed) friendship networks collected from Epinions.com ($|V| = 76K, |E| = 509K$), Slashdot.com ($|V| = 77K, |E| = 905K$), and LiveJournal.com ($|V| = 5M, |E| = 69M$) [17], denoted by Epinions, Slashdot, and LiveJournal respectively. We note that in Epinions and Slashdot, nodes are more densely connected than co-aurhorship networks, although the number of nodes for both networks are of moderate size. For the scalability test, we use synthetic power-law random networks with various sizes generated by the PYTHON Web Graph Library.

2) Propagation Probability Models: We use two propagation probability models, the Weighted cascade (WC) model and the Trivalency (TR) model which have been used as standard benchmarks in previous works so that we can compare IRIE with previous works.

- **Weighted cascade model**. Weighted cascade model [1] assigns a propagation probability to each edge by $P_{uv} = 1/d_v$ where d_v is the in-degree of v.
- **Trivalency model.** Trivalency model [4] assigns a randomly selected probability from {0.1, 0.01, 0.001} to each directed edge. This model represents the case when there several types of personal relations (three types in this case), and the edge propagation probability depends on the type of the relation.

3) Algorithms and Parameter Settings: We compare our algorithms IR and IRIE with state-of-the-art algorithms PMIA [4], CELF [5], SAEDV [8], and two baseline algorithms Degree and PageRank. Detailed parameter settings of these algorithms are described in [10].

As the stopping criteria of IR and IRIE, we use the followings. For IR and the first round of IRIE, we stop iterative computations for corresponding formulas when for all $u \in V$ difference between current $r(u)$ and the previous $r(u)$ are less than 0.0001. Otherwise the iteration run 20 rounds. For the subsequent rounds of IRIE, we initialize each $r(u)$ by the output of the previous round, and run the iterations at most 5 times and apply the same stopping criteria as in the first round.

To compare the influence spread of above algorithms, we run the Monte-Carlo simulation 10,000 times for each seed set and take the average of the influence spreads. Our experimental environment is a server with 2.8GHZ Quad-Core Intel i7 930 and 24GB memory. More simulation results including scalability test of IRIE over the network size and density, sensitivity test of IRIE to propagation probability models, and simulations on the IC-N model are presented in [10].

B. Experimental Results

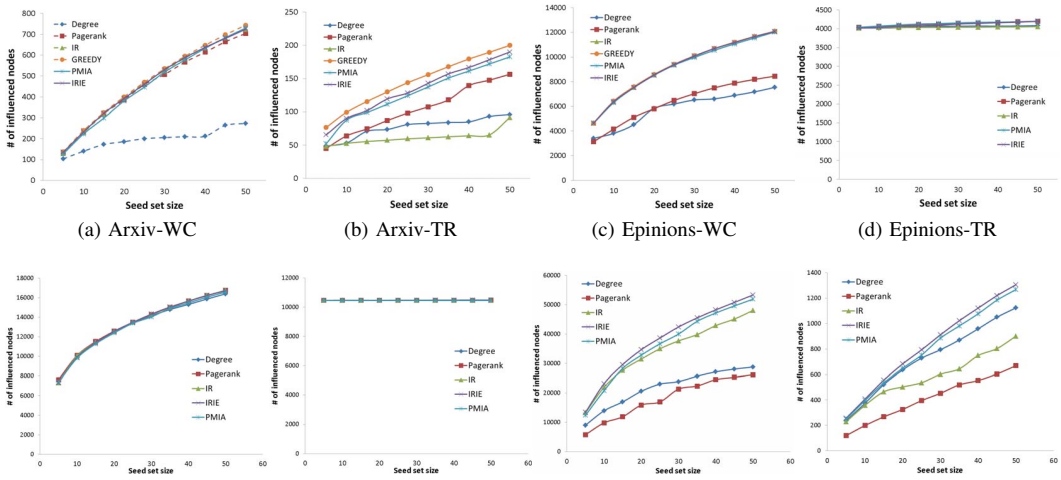
1) Influence Spread for the Real-World Datasets: We compare influence spread for each algorithms on the five real-world datasets. The seed size K is set from 1 to 50 to compare the accuracies of algorithm in various range of seed sizes. Figure IV.1 (a)-(h) shows the experimental results on influence spread. We run the CELF only for Arxiv, and Epinions(for the WC) since CELF runs too long time for other datasets.

In general, CELF performs almost the best influence spread for both the WC and the TR models. However IRIE shows almost similar performance with CELF in all cases. PMIA also shows high performance but 1-5% less influence spread than IRIE for all cases except for the Epinions TR. IR shows high performances for the WC models, but not quite good in the TR models. Hence we observe that IE part of IRIE is necessary to achieve robust performance in various steps. Hence we conclude that IRIE shows very high accuracy and robustness in most environments.

2) Running Time and Memory Usage for the Real-World Datasets: We also checked the running time of the algorithms on the real-world social networks. Figure IV.2 shows the results. The left and right figures in IV.2 corresponds to the WC model and the TR model respectively. In each figure, datasets are aligned in increasing order of network sizes from left to right. For both the WC and the TR model, IRIE is more than 1000 times faster than the CELF. Also in most cases, IRIE is quite faster than PMIA.

Note that although the numbers of nodes and edges of Epinions and Slashdot are smaller than those of DBLP, the running times of PMIA for Epinions and Slashdot are much larger than for DBLP. One possible explanation is that the running time of PMIA is sensitive to structural properties of the network such as the clustering coefficient (Epinions and Slashdot have many triangles) and edge density, and the spread size (Epinions TR and Slashdot TR induce larger spread than DBLP TR), matching the the scalability test and the sensitivity test in [10]. Hence, we conclude that IRIE shows much more stable and faster running time than PMIA in various networks.

Table I shows the experimental results on the amount of memory used by algorithms for the WC and the TR model respectively. In the table, file sizes indicate the size of raw text data files, and PMIA and IRIE indicate the amount of memory occupied by corresponding algorithms. Both for the WC model and the TR model, IRIE is 2 to 20 times more efficient in terms of memory than PMIA for all the datasets.



(a) Arxiv-WC (b) Arxiv-TR (c) Epinions-WC (d) Epinions-TR

(e) Slashdot-WC (f) Slashdot-TR (g) DBLP-WC (h) DBLP-TR

Figure IV.1: Influence spread for real world datasets

Table I: Memory usages of IRIE and PMIA

	WC			TR		
Dataset	File size	PMIA	IRIE	File size	PMIA	IRIE
ArXiv	715KB	14MB	8.7MB	582KB	10MB	8.7MB
Epinions	18MB	135MB	35MB	15MB	143MB	35MB
Slashdot	24MB	280MB	39MB	19MB	340MB	40MB
DBLP	88MB	1.1GB	160MB	82MB	357MB	158MB
LiveJournal	2.4GB	10.1GB	3GB	2GB	16GB	3GB

REFERENCES

[1] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *KDD*, 2003.

[2] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, and W. Yuan, "Influence maximization in social networks when negative opinions may emerge and propagate," in *SDM*, 2011.

[3] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *PKDD*. LNAI 4213, 2006, pp. 259–271.

[4] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *KDD*, 2010.

[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *KDD*, 2007.

[6] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks," in *WWW(Companion Volume)*, 2011.

[7] W. Chen, Y. Wang, , and S. Yang, "Efficient influence maximization in social networks," in *KDD*, 2009.

[8] Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, and K. Xie, "Simulated annealing based influence maximization in social networks," in *AAAI*, 2011.

[9] B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, and C. Faloutsos, "Threshold conditions for arbitrary cascade models on arbitrary networks," in *ICDM*, 2011.

[10] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," in *arXiv*.

[11] P. Domingos and M. Richardson, "Mining the network value of customers," in *KDD*, 2001.

[12] R. Narayanam and Y. Narahari, "A shapley value based approach to discover influential nodes in social networks," *IEEE Transactions on Automation Science and Engineering*, 2010.

[13] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *ICDM*, 2010.

[14] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, 1998.

[15] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," in *PVLDB*, 2011.

[16] Y. Wang, G. Cong, G. Song, , and K. Xie, "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks," in *KDD*, 2010.

[17] J. Leskovec, "http://snap.stanford.edu/index.html."