Maximizing the Coverage of Information Propagation in Social Networks

Zhefeng Wang[†], **Enhong Chen**[†*], **Qi Liu**[†], **Yu Yang**[‡], **Yong Ge**[§] and **Biao Chang**[†]
[†] School of Computer Science and Technology, University of Science and Technology of China
{zhefwang, chbiao}@mail.ustc.edu.cn , {cheneh, qiliuql}@ustc.edu.cn
[‡] Simon Fraser University, yya119@sfu.ca
[§] University of North Carolina at Charlotte, yong.ge@uncc.edu

Abstract

Social networks, due to their popularity, have been studied extensively these years. A rich body of these studies is related to influence maximization, which aims to select a set of seed nodes for maximizing the expected number of active nodes at the end of the process. However, the set of active nodes can not fully represent the true coverage of information propagation. A node may be informed of the information when any of its neighbours become active and try to activate it, though this node (namely informed node) is still inactive. Therefore, we need to consider both active nodes and informed nodes that are aware of the information when we study the coverage of information propagation in a network. Along this line, in this paper we propose a new problem called *Information Coverage Maximization* that aims to maximize the expected number of both active nodes and informed ones. After we prove that this problem is NP-hard and submodular in the independent cascade model, we design two algorithms to solve it. Extensive experiments on three real-world data sets demonstrate the performance of the proposed algorithms.

1 Introduction

Social network sites, such as Facebook and Twitter, have become very popular these days. These sites play important roles in the spread of information, ideas or opinions, because many people like to share their thoughts and other information on them. Thus the analysis of information propagation in social networks has been a critical research area these years.

Researchers have proposed several models to describe the diffusion of information in a social network, such as *Independent Cascade* (IC) model [Goldenberg *et al.*, 2001] and *Linear Threshold* (LT) model [Granovetter, 1978], a data-based credit distribution model [Goyal *et al.*, 2011a] and linear social influence model [Xiang *et al.*, 2013]. Among these models, IC and LT models are stochastic diffusion models [Chen *et al.*, 2013] which specify the randomized process of information propagation. In these models, each node in the network has two possible states: active and inactive. Intuitively, an active node can be viewed as adopting the new information or product that is propagated in the network.

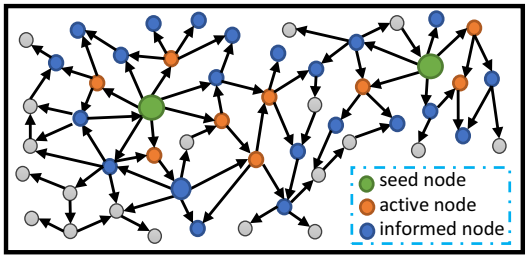


Figure 1: Information propagation in a social network

Given an information propagation model, most of the existing works focused on selecting a set of seed nodes to be activated that could lead to the maximum expected number of active nodes. This selection problem is formulated as a discrete optimization problem called *Influence Maximization* [Kempe *et al.*, 2003]. This problem, due to its important application in viral marketing, has been extensively explored ([Kimura and Saito, 2006; Wang *et al.*, 2010; Liu *et al.*, 2010; Kim *et al.*, 2013; Borgs *et al.*, 2014; Wang *et al.*, 2014]).

However, during the process of information propagation, there are actually two types of inactive nodes. For example, when we publish a message in Twitter, some users may retweet the message and others may not. But, among all users who have not retweeted the message, many of them may be aware of this message as their friends have retweeted it, while the rest is truly inactive. An example of such information propagation in a social network is shown in Figure 1. If we take a close look at the process of information propagation in this example, we will find that a node may be informed of the information if at least one of its neighbours become active. We call such nodes as *informed nodes* in this paper. In contrast, a node may never know the information if none of its neighbours is active. In fact, there are a large number of informed nodes in many real-world social networks as we will show in our experiment later. Influence maximization only considers the active nodes and neglects the informed nodes, thus it can not model the true coverage of information prop-

*Corresponding author

agation well. To better measure the coverage of information propagation, we should consider both active nodes and informed nodes.

To this end, we formulate a new problem called *Information Coverage Maximization* to address this issue. The objective of this problem is to maximize the expected number of both active nodes and informed nodes. We prove that the problem is NP-hard and submodular in the IC model. We also show that computing exact information coverage in the IC model is #P-hard. Then, we design two algorithms to solve the proposed problem. Finally, we evaluate the proposed algorithms with three real-world data sets. The experimental results demonstrate the performance of the proposed algorithms. Our contributions can be summarized as follows:

- We distinguish the informed node from the inactive node, and explore the value of informed nodes to better measure the coverage of information propagation. Thus, we propose a new problem of maximizing the expected number of both active nodes and informed nodes.
- We prove that the proposed problem is NP-hard and submodular in the IC model. We also show the computation of information coverage in the IC model is #P-hard.
- We design two algorithms to solve the proposed problem. The proposed algorithms are examined with three real-world data sets and the experimental results show the performance of the proposed algorithms.

2 Related Work

Social networks have been studied extensively for many years. A rich body of these studies is focused on the analysis of influence and information propagation in social networks. Several models have been proposed to describe the diffusion of information through the social network, such as IC model [Goldenberg *et al.*, 2001], LT model [Granovetter, 1978] and decreasing cascade model [Kempe *et al.*, 2005]. These models define the stochastic process of information propagation. Thus they are called stochastic diffusion models [Chen *et al.*, 2013]. There are also models which formulate the information propagation from other perspectives ([Aggarwal *et al.*, 2011; Goyal *et al.*, 2011a; Xiang *et al.*, 2013]). Moreover, in [Chen *et al.*, 2012] and [Liu *et al.*, 2012], the authors extended IC model to consider the time-delay aspect of influence diffusion.

Influence maximization [Kempe *et al.*, 2003], which aims to maximize the expected number of active nodes in a given diffusion model, is another main research direction of the analysis of information propagation in social networks. In [Kempe *et al.*, 2003], the authors proved the problem is NP-hard in both IC and LT models and proposed a greedy framework to solve it. The following researchers focused on developing both efficient and effective algorithms, such as CELF [Leskovec *et al.*, 2007], PMIA [Chen *et al.*, 2010a], LDAG [Chen *et al.*, 2010b], SIMPATH [Goyal *et al.*, 2011b], StaticGreedy [Cheng *et al.*, 2013], Linear and Bound [Liu *et al.*, 2014] and IMRank [Cheng *et al.*, 2014]. In addition, in [Chen *et al.*, 2012] and [Liu *et al.*, 2012], the authors studied the influence maximization with time-critical constraint. In [Tang *et al.*, 2014], the authors studied the diversified influence maximization which considers both the magnitude of influence and the diversity of the influenced crowd. But influence maximization only considers the active nodes, which makes it different from the proposed problem.

3 Problem Formulation

In this section, we first give a formal statement of information coverage maximization problem. Then we discuss the properties of the proposed problem.

3.1 Problem Definition

Let the directed graph $G = (V, E, T)$ denote an information propagation network, where $V = \{1, 2, ..., n\}$ is the set of nodes in the graph and E denotes all edges that include all the information propagation paths between nodes. And $T = [t_{i,j}]_{n*n}$ is the propagation probability matrix, where $t_{i,j}$ represents the probability of information propagation from node i to node j. We use n to denote the number of nodes and m to denote the number of edges respectively.

Although we can use different stochastic diffusion model to describe the information propagation process, we adopt IC model in this paper as it has been shown as one of the most suitable models for the diffusion of information [Chen *et al.*, 2013]. In the model, seed nodes are the initial nodes selected to propagate the information and they will try to activate their neighbours. Their neighbours will be informed in this process and may be activated. If a node is activated, it becomes an active node and will try to activate its own neighbours. If a node is not activated but receives the information, then it is an informed node. The process converges when no more nodes can be activated.

Let S, A and L denote the seed nodes, active nodes and informed nodes respectively. Then we can get the relationship between A and S as follows:

$$A = I(S) \tag{1}$$

where $I(S)$ are the active nodes when the information diffusion process converges. Similarly, we can get the relationship between L and A as follows:

$$L = \bigcup_{a \in A} N(a) \tag{2}$$

where $N(a)$ are the inactive out neighbours of node a. To this end, we can formulate the *Information Coverage Maximization Problem* as follows:

$$\begin{aligned} \arg\max_{S} \ & F(S) = E(|A|) + E(|L|) \\ s.t. \ & |S| = k \end{aligned} \tag{3}$$

where k is a given budget. $E(\cdot)$ is the expectation of the number, because information propagation is a stochastic process. $F(S)$ is the sum of the expected number of active nodes and informed nodes. We will refer to it as *information coverage* in this paper.

Considering the relationship given by Eq.(1) and Eq.(2), we can rewrite the Eq.(3) as follows:

$$\arg\max_{S} \; F(S) = E(|I(S)|) + E(|\bigcup_{a \in I(S)} N(a)|) \quad s.t. \quad |S| = k \tag{4}$$

Comparing Eq.(4) to the objective function of the traditional influence maximization problem, we can find that the first term of Eq.(4) is exactly the influence spread of S. But Eq.(4) has an extra term $E(|\bigcup_{a \in I(S)} N(a)|)$ which is the expected number of informed nodes. Thus information coverage maximization can better model the true coverage of information propagation in a social network.

In the real world, the informed nodes may have different values than the active nodes. Thus we introduce a weight to adjust the contribution of informed node to the measure of information coverage. Along this line, we propose a general form of information coverage maximization problem: *Weighted Information Coverage Maximization Problem.* We formulate it as follows:

$$\arg\max_{S} \; W(S) = E(|I(S)|) + \lambda \, E(|\bigcup_{a \in I(S)} N(a)|) \quad s.t. \quad |S| = k \quad and \quad \lambda \in [0, 1] \tag{5}$$

where λ is the weight coefficient that controls the importance of informed nodes. When λ equals 0, the problem degenerates into the traditional influence maximization problem. When λ equals 1, the problem is the same as the information coverage maximization problem.

3.2 Problem Property

In this part, to show the properties of the problem, we prove several theorems about the problem.

Theorem 1 *For an information propagation network formulated by IC model, the information coverage maximization problem is NP-hard.*

Proof. Consider the NP-complete set cover problem [Karp, 1972]: given a collection of subsets $S_1, S_2, ..., S_m$ of a ground set $U = \{u_1, u_2, ..., u_n\}$; the question is if there exist k of the subsets whose union equals to U. We will reduce the problem to the information coverage maximization problem.

Given an arbitrary instance of the set cover problem, we construct a corresponding directed bipartite graph: there is a node i for each subset S_i, a node j for each element u_j, and a directed edge(i, j) with a propagation probability $t_{i,j} = 0$ when $u_j \in S_i$. Since all probabilities are 0, the propagation is a deterministic process. Thus, the set cover problem is equivalent to deciding if there is a set N of k nodes in the graph with $F(N) = n + k$. If any set N of k nodes has $F(N) = n + k$, then we can initially activate the k nodes corresponding to subsets such that all n nodes corresponding to elements in the ground set will be informed. This means that the set cover problem must be solvable.

Theorem 2 *For an information propagation network formulated by IC model, the weighted information coverage maximization problem is NP-hard.*

Proof. Since the information coverage maximization problem is a special case of the weighted information coverage maximization problem, the result is straightforward.

Theorem 3 *For an information propagation network formulated by IC model, computing the information coverage $F(S)$ or the weighted information coverage $W(S)$ is #P-hard.*

Proof. Consider the #P-complete $s-t$ connectedness problem [Valiant, 1979]: given a directed graph $G = (V, E)$ and two nodes s and t in the graph; the question is to count the number of subgraphs of G in which s is connected to t. It is straightforward to see that this problem is equivalent to computing the probability that s is connected to t when each edge in G is connected with a probability of $\frac{1}{2}$.

Given an arbitrary instance of the $s-t$ connectedness problem, we will reduce it to the computation of information coverage. let $F_G(S)$ denote the information coverage of seed set S in graph G. Then let $S = \{s\}$ and $p(e) = \frac{1}{2}$ for all $e \in E$, and compute $I_1 = F_G(S)$. Next, add a new node t' and a directed edge from t to t' with propagation probability $p_{t,t'} = 1$. Now we obtain a new graph G' and compute $I_2 = F_{G'}(S)$. Let $p_G(S, t)$ denote the probability that node t is activated by S. Since graph G' has an extra node t', it is easy to see that $I_2 = F_G(S) + p_G(S, t)(p_{t,t'} + 1 - p_{t,t'})$. Thus, $I_2 - I_1$ is the probability that s is connected to t. This means that $s - t$ connectedness problem must be solvable. For the $W(S)$ case, replace $F(\cdot)$ with $W(\cdot)$ and $I_2 = W_G(S) + p_G(S, t)(p_{t,t'} + \lambda(1 - p_{t,t'}))$.

Theorem 4 *For an information propagation network formulated by IC model, $F(\cdot)$ is monotone and submodular.*

Proof. Since the monotonicity of $F(\cdot)$ is straightforward, we focus on proving its submodularity. Given a graph $G = (V, E, T)$, we can construct live-arc graphs for the IC model according to the methods proposed in [Kempe *et al.*, 2003]. Then let G_L be a random live-arc graph, and let the $Prob(G_L)$ denote the probability that G_L is selected from all possible live-arc graphs. Let $R_{G_L}(S)$ denote the set of all nodes that can be reached from S in G_L. For traditional influence maximization problem, the expected numbers of $R_{G_L}(S)$ is exactly the influence spread of S. However, in our case, we need to add the inactive out neighbour nodes of the active nodes. Then we use $Q_{G_L}(S)$ to denote the union of the $R_{G_L}(S)$ and the out neighbour nodes of $R_{G_L}(S)$ in G. Thus for IC model, we have

$$F(S) = \sum_{all \; possible \; G_L} Prob(G_L)|Q_{G_L}(S)| \tag{6}$$

Since an non-negative linear combination of submodular functions is also submodular, we just need to prove $|Q_{G_L}(\cdot)|$ is submodular for any live-arc graph. To do this, Let M and N be two sets of nodes such that $M \subseteq N$, and consider the number $|Q_{G_L}(M \cup v)| - |Q_{G_L}(M)|$. This is the number of elements in $Q_{G_L}(v)$ that are not already in the $Q_{G_L}(M)$. Thus it must be greater or equal to the number of elements in $Q_{G_L}(v)$ that are not already in the $Q_{G_L}(N)$. It follows that $|Q_{G_L}(M \cup v)| - |Q_{G_L}(M)| \geq |Q_{G_L}(N \cup v)| - |Q_{G_L}(N)|$. Thus function $|Q_{G_L}(\cdot)|$ is submodular, which means that $F(\cdot)$ is submodular.

Theorem 5 *For an information propagation network formulated by IC model, $W(\cdot)$ is monotone and submodular.*

Proof sketch. We can utilize the live-arc graph technique to prove this theorem in a similar way as the proof of Theorem 4.

4 Solution

In this section, we propose two algorithms to solve the problem. First, we discuss a greedy algorithm with "Lazy forward" update scheme. Second, we discuss a degree-based heuristic algorithm.

We have proved that both $F(\cdot)$ and $W(\cdot)$ are monotone and submodular, and they apparently satisfy $F(\emptyset) = 0$ and $W(\emptyset) = 0$. Thus, according to [Nemhauser *et al.*, 1978], a simple greedy algorithm can approximate the optimal solution with a factor of $1 - 1/e$. Since we have proved that computing $F(\cdot)$ or $W(\cdot)$ is #P-hard, we have to run Monte Carlo simulations for sufficiently many times (e.g., 10,000) to estimate the value of $F(\cdot)$ or $W(\cdot)$. Consequently, the simple greedy algorithm is very time-consuming. Inspired by [Leskovec *et al.*, 2007], we also design a "Lazy Forward" update scheme for our algorithm. Due to the submodularity of the problem, this update scheme can reduce the times of estimating $F(\cdot)$ or $W(\cdot)$. More details about the update scheme are shown in Algorithm 1. From the algorithm, we can see that it needs $(n + k\beta)$ times of information coverage estimations, where $\beta \ll n$ is the expected number of information coverage estimations in each iteration. Thus the total time cost is $O(nRm + k\beta Rm)$, where R is the number of rounds of simulations in each estimation.

Algorithm 1: The Lazy-Forward Greedy Algorithm

```
Input: G = (V, E, T), number k
Output: seed set S
initialize S = ∅
for each node n in V do
    //for the weighted case, replace F(·) with W(·)
    compute Δ(n) = F(n)
    stamp_n = 0
end
while |S| < k do
    n = arg max_{n∈V\S} Δ(n)
    if stamp_n == |S| then
        S = S ∪ n
    end
    else
        //for the weighted case, replace F(·) with W(·)
        compute Δ(n) = F(S ∪ n) − F(S)
        stamp_n = |S|
    end
end
return S
```

Although "Lazy Forward" update scheme reduces the time cost dramatically, it is still intractable for large scale networks. In the real world, there are often thousands of nodes and millions of edges in a social network. To address the scalability issue, we develop an efficient heuristic algorithm. When we revisit the objective function, we can find that a node's contribution to the information coverage is highly dependent on its out degree. Thus if we rank the nodes according to their out degrees and take top-k nodes as the seed nodes, we can probably get a good result. Furthermore, when a node is selected, its out neighbours will be informed. This will result in a decrease of other nodes' "effective" out degrees, as their out neighbours may have been informed. This observation means that we can benefit from adjusting each node's "effective" out degree dynamically. This heuristic is summarized in Algorithm 2. From the algorithm, we can see that it takes only $O(k(n + m))$ time to complete if we store the graph G and the covered nodes set C with appropriate data structures.

Algorithm 2: The Effective Degree Rank Algorithm

```
Input: G = (V, E, T), number k
Output: seed set S
initialize S = ∅
initialize C = ∅
for each node n in V do
    EffectiveDegree(n) = OutDegree(n)
end
while |S| < k do
    n = arg max_{n∈V\S} EffectiveDegree(n)
    S = S ∪ n
    C = C ∪ OutNeighbour(n)
    for each node n in V \ S do
        EffectiveDegree(n) =
        OutDegree(n) − |C ∩ OutNeighbour(n)|
    end
end
return S
```

5 Experiment

In this section, we explore the difference between influence maximization and information coverage maximization. Then we demonstrate the performance of the proposed algorithms.

5.1 Experimental Setup

Data Sets. The three real-world data sets we used are: **wiki-Vote** which is the Wikipedia who-votes-on-whom network, **soc-Epinions1** which is the who-trusts-whom network of Epinions.com [1], and **weibo** which is the who-follows-whom network of Weibo.com [2]. The first two are downloaded from SNAP[3], and the last one is crawled from Weibo.com, which is a Chinese microblogging site like Twitter. More detailed information about the data sets is shown in Table 1.

The social networks are constructed like this: If a node j votes (trusts or follows) another node i, there is a directed edge from node i to node j. The propagation probability of an

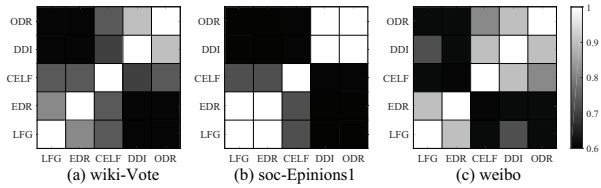
[1] http://www.epinions.com/

[2] http://weibo.com/

[3] http://snap.stanford.edu

Table 1: Statistics of data sets

Data set	Type	#Nodes	#Edges
wiki-Vote	Directed	7,115	103,689
soc-Epinions1	Directed	75,879	508,837
weibo	Directed	76,491	9,572,897



(a) wiki-Vote (b) soc-Epinions1 (c) weibo

Figure 2: The Jaccard similarity coefficient of selected nodes

edge (i, j) is set to be $\frac{weight(i,j)}{indegree(j)}$, as widely used in literatures ([Chen *et al.*, 2009; Goyal *et al.*, 2011b]).

Algorithms for Comparison. We compare the proposed algorithms with several influence maximization algorithms. The algorithms we used in the experiments include:

- *LFG* is the Algorithm 1 proposed in section 4.
- *EDR* is the Algorithm 2 proposed in section 4.
- *CELF* is an approximation algorithm for influence maximization proposed in [Leskovec *et al.*, 2007].
- *DegreeDiscountIC (DDI)* is a degree-based heuristic algorithm for influence maximization. We set the parameter p to 0.01, same as used in [Chen *et al.*, 2009].
- *OutDegreeRank (ODR)* outputs the top-k nodes with the highest out degree.

Evaluation Method. With the output of each algorithm, we use it as the seed set to compute their information coverage in the IC model. In the computation process, we run Monte Carlo simulation $10,000$ times to obtain an estimation of the information coverage.

We implemented the algorithms in Java and conducted the following experiments on a Linux server with two 2.0GHz Six-Core Intel Xeon E5-2620 and 96G memory.

5.2 Experimental Results

Correlation Demonstration. For the purpose of demonstrating the difference between influence maximization and information coverage maximization, Figure 2 shows the the Jaccard similarity coefficient of the seed sets selected by different algorithms when the size of seed set is 20. In Figure 2, we can see that the seed sets selected by *LFG* and *EDR* for information coverage maximization are similar to each other, while *CELF* for influence maximization selects different seed nodes from them. Meanwhile, *DDI* and *ODR* select similar seed nodes and these nodes are more similar to the ones selected by *CELF* than *LFG* and *EDR*. This phenomenon further shows the difference between influence maximization and information coverage maximization.

Effectiveness validation. We run tests on three social networks to obtain information coverage results. The size of seed set ranges from 5 to 20. Figure 3 (a), (b) and (c) show the information coverage results on three social networks. From the figure, we can see that *LFG* and *EDR* obviously outperform the other algorithms. This result proves once more the difference between influence maximization and information coverage maximization. Meanwhile, we can find that *LFG* and *EDR* have similar performance. This result demonstrates the effectiveness of "effective" degree heuristic. Furthermore, *CELF* has better performance than *DDI* and *ODR*, which means that information diffusion still plays an important role in the result. For better illustration, Figure 3 (d) shows the comparative results of different algorithms in weibo data set, where the performance of *CELF* is chosen as the baseline and the size of the seed set ranges from 10 to 20. From the Figure, we can see that *LFG* has the best performance. Also, *EDR* performs better than *CELF* significantly. Finally, *DDI* and *ODR* perform worse than *CELF* in most cases.

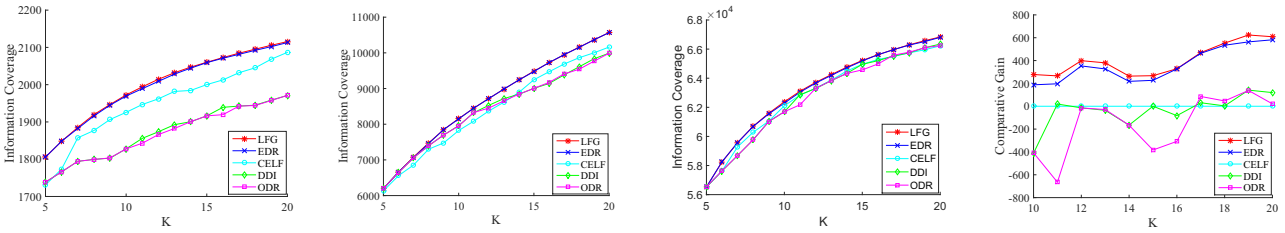
Table 2: Number of nodes of different types

Data set	#Active	#Informed	#Other
wiki-Vote	42	2,073	5,000
soc-Epinions1	274	10,259	65,346
weibo	465	66,362	9,674

We also run tests to verify the weighted information coverage case. We set λ to 0.5 and 0.25 respectively. Figure 4 shows the weighted information coverage results on the same three social networks. From the figure, we can find that *LFG* and *EDR* still have better performance than the other algorithms. But the gap between *CELF* and *LFG* does become smaller, which means that information diffusion becomes more important when the relative values of informed nodes become smaller. We also show the comparative results of different algorithms in Figure 4 (d) and (h). From the figure, we can see that the result is similar to the one shown in Figure 3 (d).

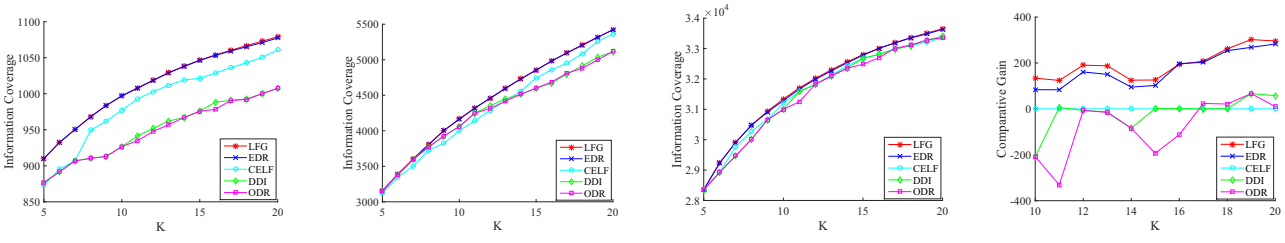
Case study. We run Monte Carlo simulations to obtain the number of nodes of different types (e.g., active nodes) after we get the seed set selected by *LFG*. Table 2 shows the statistic result when the size of seed set is 20. From the table, we can find that the number of active nodes is very small while the number of informed nodes is large. This phenomenon is especially significant in weibo data set. The reason is that nodes in weibo data set have more neighbours on average. Furthermore, we calculate the ratio of nodes of different types. Results are shown in Figure 5. From the figure, we can see that the total ratio of active nodes and informed nodes is considerable even if there are only 20 seed nodes. It means that we only need a small number of seed nodes if we want to get a large information coverage.

Efficiency Comparison. Table 3 shows the running time of different algorithms when the size of seed set is 20. From the table, we can see that *DDI* and *ODR* are more efficient than the other algorithms. Also, *EDR* has a very low time cost while *LFG* and *CELF* are quite time-consuming. The reason is that we need to run Monte Carlo simulations to estimate information coverage and influence spread in *LFG* and *CELF* respectively. Remarkably, to improve the efficiency of the two algorithms, we have already utilized Java multi-thread



(a) Wiki-Vote (b) soc-Epinions1 (c) weibo (d) weibo

Figure 3: Information coverage on three social networks.



(a) Wiki-Vote, $\lambda = 0.5$ (b) soc-Epinions1, $\lambda = 0.5$ (c) weibo, $\lambda = 0.5$ (d) weibo, $\lambda = 0.5$

(e) Wiki-Vote, $\lambda = 0.25$ (f) soc-Epinions1, $\lambda = 0.25$ (g) weibo, $\lambda = 0.25$ (h) weibo, $\lambda = 0.25$

Figure 4: Weighted information coverage on three social networks

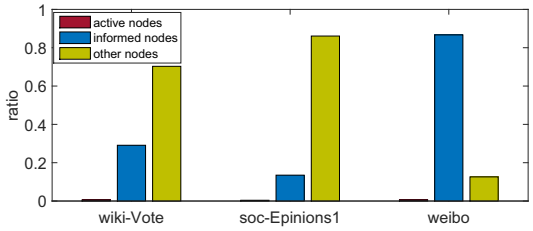


Figure 5: Node type distribution on three social networks

technology to speed up the simulation process.

6 Conclusion

In this paper, to better measure the coverage of information propagation, we distinguish the informed node from the inactive node and explore the value of the informed nodes. Meanwhile, we formulate a novel problem called information coverage maximization which aims to maximize the expected number of both active nodes and informed nodes. Furthermore, we prove the proposed problem is NP-hard and submodular in the IC model. We also show that the computation of information coverage in the IC model is #P-hard. Then based on the properties of the problem, we design two algorithms to solve it. Finally, we conduct extensive experiments to verify our idea. The experimental results show the difference between influence maximization and information coverage maximization. The performance of the proposed algorithms is also demonstrated in the experiments. We hope our study could lead to more future works.

Table 3: Efficiency comparison (in seconds)

Data set	LFG	EDR	CELF	DDI	ODR
wiki-Vote	881.4	0.268	15.19	0.014	0.02
soc-Epinions1	940.6	1.169	63.36	0.025	0.066
weibo	6.6e5	23.78	866.6	0.121	0.086

7 Acknowledgments

This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the Natural Science Foundation of China (Grant No. 61403358), the Fundamental Research Funds for the Central Universities of China (Grant No. WK0110000042), the Anhui Provincial Natural Science Foundation (Grant No. 1408085QF110), and the National

Center for International Joint Research on E-Business Information Processing under Grant 2013B01035.

References

[Aggarwal *et al.*, 2011] Charu C Aggarwal, Arijit Khan, and Xifeng Yan. On flow authority discovery in social networks. In *SDM*, pages 522–533. SIAM, 2011.

[Borgs *et al.*, 2014] Christian Borgs, Michael Brautbar, Jennifer T Chayes, and Brendan Lucier. Maximizing social influence in nearly optimal time. In *SODA*, pages 946–957. SIAM, 2014.

[Chen *et al.*, 2009] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *SIGKDD*, pages 199–208. ACM, 2009.

[Chen *et al.*, 2010a] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *SIGKDD*, pages 1029–1038. ACM, 2010.

[Chen *et al.*, 2010b] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *ICDM*, pages 88–97. IEEE, 2010.

[Chen *et al.*, 2012] Wei Chen, Wei Lu, and Ning Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In *AAAI*, 2012.

[Chen *et al.*, 2013] W. Chen, Laks V.S. Lakshmanan, and C. Castillo. *Information and Influence Propagation in Social Networks*. Morgan and Claypool, 2013.

[Cheng *et al.*, 2013] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In *CIKM*, pages 509–518. ACM, 2013.

[Cheng *et al.*, 2014] Suqi Cheng, Huawei Shen, Junming Huang, Wei Chen, and Xueqi Cheng. Imrank: Influence maximization via finding self-consistent ranking. In *SIGIR*, pages 475–484. ACM, 2014.

[Goldenberg *et al.*, 2001] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223, 2001.

[Goyal *et al.*, 2011a] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 5(1):73–84, 2011.

[Goyal *et al.*, 2011b] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *ICDM*, pages 211–220. IEEE, 2011.

[Granovetter, 1978] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420, 1978.

[Karp, 1972] Richard M Karp. *Reducibility among combinatorial problems*. Springer, 1972.

[Kempe *et al.*, 2003] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146. ACM, 2003.

[Kempe *et al.*, 2005] David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Automata, languages and programming*, pages 1127–1138. Springer, 2005.

[Kim *et al.*, 2013] Jinha Kim, Seung-Keol Kim, and Hwanjo Yu. Scalable and parallelizable processing of influence maximization for large-scale social networks? In *ICDE*, pages 266–277. IEEE, 2013.

[Kimura and Saito, 2006] Masahiro Kimura and Kazumi Saito. Tractable models for information diffusion in social networks. In *Knowledge Discovery in Databases: PKDD 2006*, pages 259–271. Springer, 2006.

[Leskovec *et al.*, 2007] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *SIGKDD*, pages 420–429. ACM, 2007.

[Liu *et al.*, 2010] Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. Mining topic-level influence in heterogeneous networks. In *CIKM*, pages 199–208. ACM, 2010.

[Liu *et al.*, 2012] Bo Liu, Gao Cong, Dong Xu, and Yifeng Zeng. Time constrained influence maximization in social networks. In *ICDM*, pages 439–448, 2012.

[Liu *et al.*, 2014] Qi Liu, Biao Xiang, Enhong Chen, Hui Xiong, Fangshuang Tang, and Jeffrey Xu Yu. Influence maximization over large-scale social networks: A bounded linear approach. In *CIKM*, pages 171–180. ACM, 2014.

[Nemhauser *et al.*, 1978] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[Tang *et al.*, 2014] Fangshuang Tang, Qi Liu, Hengshu Zhu, Enhong Chen, and Feida Zhu. Diversified social influence maximization. In *ASONAM*, pages 455–459. IEEE, 2014.

[Valiant, 1979] Leslie G Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[Wang *et al.*, 2010] Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *SIGKDD*, pages 1039–1048. ACM, 2010.

[Wang *et al.*, 2014] Zhefeng Wang, Hao Wang, Qi Liu, and Enhong Chen. Influential nodes selection: a data reconstruction perspective. In *SIGIR*, pages 879–882. ACM, 2014.

[Xiang *et al.*, 2013] Biao Xiang, Qi Liu, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. Pagerank with priors: An influence propagation perspective. In *IJCAI*, 2013.