Holistic Influence Maximization: Combining Scalability and Efficiency with Opinion-Aware Models

Sainyam Galhotra*
College of Computer Science
UMass Amherst
sainyam@cs.umass.edu

Akhil Arora*
Text & Graph Analytics
XRCI, Bangalore
akhil.arora@xerox.com

Shourya Roy
Text & Graph Analytics
XRCI, Bangalore
shourya.roy@xerox.com

ABSTRACT

The steady growth of graph data from social networks has resulted in wide-spread research in finding solutions to the influence maximization problem. In this paper, we propose a *holistic* solution to the influence maximization (*IM*) problem. (1) We introduce an *opinion-cum-interaction* (OI) model that closely mirrors the real-world scenarios. Under the OI model, we introduce a novel problem of *Maximizing the Effective Opinion (MEO)* of influenced users. We prove that the *MEO* problem is NP-hard and cannot be approximated within a constant ratio. (2) We propose a heuristic algorithm *OSIM* to efficiently solve the MEO problem. To better explain the *OSIM* heuristic, we first introduce *EaSyIM* – the opinion-oblivious version of *OSIM*, a scalable algorithm capable of running within practical compute times on commodity hardware. In addition to serving as a fundamental building block for *OSIM*, *EaSyIM* is capable of addressing the scalability aspect – *memory consumption* and *running time*, of the IM problem both theoretically and empirically.

1. INTRODUCTION

Social networks have become pervasive owing to the exponential growth in their popularity. The sheer scale at which these networks operate today is humongous. Thus, the *influence maximization* (IM) problem [4] with its applicability in solving a host of real-world problems and beyond, has been one of the most widely studied problems over the past decade. This problem is to identify a set of *seed nodes* so that the overall *spread* of information in a social network, which is the potential collective impact of imparting that piece of information to these nodes, is maximized.

Although widely studied from the aspect of designing efficient and scalable algorithms [3–5] for IM, research on devising novel information diffusion models capable of mirroring the dynamics of information propagation under real-world scenarios has been negligible. The most obvious limitation of the funadmental models is that the notion of *spread*, of a set of seed-nodes S, is defined as a function of the number of nodes that get activated using these seeds. The problem with this definition is two fold – (1) Each node is considered to be contributing fully and positively towards the *spread* of information about a content without considering the personal *opinion* of the nodes which could be even negative, (2) A newly active node is always considered to perceive the information with the same intent as that of the node that activated it, while the former may tend to disagree with the latter owing to the *interactions* between them in the past.

Since these scenarios are not scarce in the real-world settings, in this paper, we propose a new *Opinion-cum-Interaction* (OI) model of information diffusion capable of addressing the limitations discussed above and introduce a more realistic notion of *opinion-spread*. Under the OI model, we propose a novel problem (MEO) of maximizing the *opinion-spread*. Since *MEO* is NP-hard and difficult to approximate within a constant ratio, we incorporate ideas from the *opinion-oblivious* scenarios to design *scalable* algorithms for the *opinion-aware* case. Having analyzed the state-of-the-art, it was evident that algorithms that are *efficient* are not *scalable* and vice-versa. We propose *OSIM* (opinion-aware) and *EaSyIM* (opinion-oblivious) algorithms, that run in *linear* time and space thus, providing the best tradeoff between *memory-consumption* and *running-time* and the capability of handling real-world large scale networks on moderately sized machines. For details, we kindly refer the reader to the full version [2] of this work[1].

2. MODEL AND ALGORITHM

We first describe the OI model, which serves as an extension over the IC and LT models with modifications to include a second layer attributed to modelling the diffusion and change of *opinion* of the activated users; to facilitate *opinion-aware* IM. As opposed to the IC/LT models, where a newly activated node (oblivious to its *opinion*) is always considered to be contributing positively towards the information spread, the OI model considers the *spread* (opinion-spread) of information under an *opinion-aware* scenario – where the contribution of a newly activated node could as well be negative. OI makes use of *opinion* of a node and *interaction* between two nodes to simulate the opinionated flow of information over the network. The *opinion* of a node, o is defined as the orientation of the node coupled with its strength signifying the preference towards a content, thus, $o \in [-1, 1]$. The *interaction* probability between two nodes u, v is defined as a fraction of the times a content shared by u gets accepted by v with the same orientation as that of u.

The diffusion dynamics of OI are as follows. The first step is the activation step which is the same as that of IC and LT models. This is followed by an opinion estimation step, where the opinion of an activated node is derived using its personal opinion and the opinion of the node(s) that activate(s) it. More formally, the contribution of the node u, which activates the other node v, is o_u with a probability equal to the interaction between them and $-o_u$ otherwise. In this way, opinions are assigned to the activated nodes in the network and the information diffusion stops when no new node gets activated. Seeds are selected based on the *opinion-spread*, which is defined as the sum of the updated opinions of all the activated nodes.

As mentioned in Sec. 1, in order to scalably solve the MEO problem, we propose *OSIM*, capable of maximizing the *opinion-spread* under the OI model. With the same fundamental ideas, we propose *EaSyIM* for the opinion-oblivious settings. Owing to space constraints, next, we describe the latter alone in detail.

*The first two authors have contributed equally to this work.

[1] https://people.cs.umass.edu/~sainyam/SIGMOD16.pdf

Algorithm 1 Seed Selection using EaSyIM

```
Input: Graph $G = (V, E)$, #seeds ($k = |S|$) and path-length ($l$)
Output: Seed set $S$
$S,\ C \leftarrow \emptyset$
for i = 1 to k do
    $max,\ maxId \leftarrow 0$
    $Score \leftarrow AssignScore(G, C, l)$
    for each $u \in V$ do
        if $Score[u] > max$ then
            $max \leftarrow score; maxId \leftarrow u$
        end if
    end for
    $S \leftarrow S \cup \{maxId\}; C \leftarrow C \cup \mathcal{F}(maxId)$
end for
```

Our algorithms take a graph G, the number of seeds k and a path-length value l as input and returns a set of seed nodes S. The fundamental building block of our algorithms is a *score-assignment* step, which leverages the idea that the probability of a node v to get activated by a seed node u is dependent upon the number of simple paths from u to v in G. Thus, a simple function of the number of paths from a node u to all other nodes $v \in V \setminus \{u\}$ can be used to assign a score to u. This score-assignment is further used to rank all the nodes $v \in V$ in an order determining their expected spread $\sigma(v)$. The following explanations of the score assignment algorithms assume the IC model of information diffusion. Their extensions to the *linear threshold* (LT) and the *weighted cascade* (WC) models are present in [2].

Next, we describe the score assignment step for the opinion-oblivious (EaSyIM) case, and refer the reader to [2] for a very similar description of *OSIM*. The scores assigned to a node u ($Score[u]$), is computed by aggregating the contributions of all $u \rightsquigarrow v$ paths of length at most l^2 ($\mathcal{L}(u \rightsquigarrow v) \leq l$). The weight for each path is defined as the product of probabilities $p(e)$ of the edges composing that path. Both score-assignment algorithms, try to mimic closely the expected value of the spread.

At each iteration, our algorithm selects the node with the maximum score as the seed node and updates the set C with the nodes activated by this seed. At any given iteration, the set C contains all the nodes activated $\mathcal{F}(s)$ by the seed nodes $s \in S$. In order to ensure that the set of nodes activated by each selected seed node are disjoint, $AssignScore$ neglects the contribution of all the paths containing any previously activated node $c \in C$ in the score calculation for the subsequent iterations.

Paths of length l from a node u can be calculated as the sum of all paths of length $l - 1$ from its neighbors. Owing to this observation, the time taken by the algorithm to assign scores to each node of the graph is $O(l(m+n))$ because for each iteration over l, it looks at the adjacency list of each node to updated the score. Hence the overall time complexity of *EaSyIM*, and similarly *OSIM*, is $O(kl(m+n))$.

3. EXPERIMENTS

All the simulations were done using the Boost graph library in C++ on an Intel(R) Xeon(R) 20-core machine with 2.4 GHz CPU and 100 GB RAM running Linux Ubuntu 12.04. We present results on real (large) graphs (details in [1, 2]), taken from the arXiv and SNAP repositories. We adopt the C++ implementation of CELF++ and TIM$^+$ made available by the respective authors. We consider a mix of both directed and undirected graphs, however to ensure uniformity the undirected graphs were made directed by considering, for each edge, the arcs in both the directions. As a conventional practice, the *spread*, and hence the *opinion-spread* as well, is calculated as an average over $10K$ Monte Carlo (MC) simulations[3].

Dataset	Running Time (min)			Memory (MB)		
	CELF++	EaSyIM (l=1)	Gain	CELF++	EaSyIM (l=1)	Gain
NetHEPT	5352.25	118	**45.35x**	23.26	3.39	6.86x
HepPh	9746.74	230	**41x**	24.60	3.47	7.08x
DBLP	NA	5071.67	**∞**	NA	44.73	**∞**

Table 1: **Comparing *EaSyIM* with CELF++,** $k = 100$ **and** $l = 1$**.**

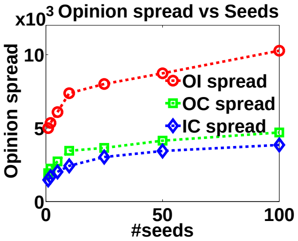

Figure 1: **Opinion-spread** ($\lambda = 1$) **comparison of OI with OC and IC on Twitter data.**

The first set of results on the data extracted from the *Twitter* social network, portray the importance of the *OI* model in real-world scenarios over the legacy information diffusion models, namely – IC, WC and LT and the state-of-the-art OC model in the context of opinion-aware IM [6]. Figure 1 shows that the *opinion-spread* achieved using the seeds selected by the OI model is much better when compared to that of the OC and the IC model.

In addition to the results on *quality*, we next portray the *scalability* of our algorithms. It is evident from Tables 1 and 2 that, *EaSyIM* is 10–15 times more efficient while consuming 3-4 times less memory when compared to CELF++, and requires 8–10 times more time to run while its memory-footprint is ≈ 500 times smaller when compared to TIM$^+$. We argue that since *EaSyIM* can be efficiently parallelized owing to the independence of the MC simulations, its lack of efficiency when compared to TIM$^+$ can be easily mitigated by running it in parallel on 8 cores[4] while ensuring the memory gain to be the same. Note that, the memory consumed by *TIM*$^+$ was $>$ 100GB for $\epsilon = 0.1$, $k = 50$ both on YouTube and socLive, thus, we could not report the results for the same.

4. CONCLUSIONS

In this paper, we addressed the problem of influence maximization in social networks under a very generic opinion-aware setting, where the nodes can possess any one of the – positive, neutral and negative opinions. To this end, we introduced the novel *MEO* problem and devised a *holistic* solution to the influence maximization problem; by coming up with an opinion-cum-interaction (*OI*) model, and scalable algorithms – *OSIM* and *EaSyIM*. Since majority of the works in the literature operate oblivious to the existence of opinions and are not scalable, there are efficiency concerns in real-world scenarios with huge graphs. Consequently, we designed efficient algorithms that run in time and space linear to the size of the graph; which is orders of magnitude better when compared to the state-of-the-art techniques. Our empirical studies on real-world social network datasets showed that our algorithms are effective, efficient, scale well – providing the best trade-off between running time and memory consumption, and are practical for large real graphs. In future, we would like to come up with distributed versions of our algorithms, thus enabling them scale to even larger graphs.

5. REFERENCES

[1] S. Galhotra, A. Arora, and al. Asim: A scalable algorithm for influence maximization under the ic model. In *WWW (Companion Volume)*, 2015.

[2] S. Galhotra, A. Arora, and S. Roy. Holistic influence maximization: Combining scalability and efficiency with opinion-aware models. In *SIGMOD*, 2016.

[3] A. Goyal and al. Celf++: Optimizing the greedy algorithm for influence maximization in social networks. In *WWW (Companion Volume)*, 2011.

[4] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[5] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. In *SIGMOD*, pages 75–86, 2014.

[6] H. Zhang, T. N. Dinh, and M. T. Thai. Maximizing the spread of positive influence in online social networks. In *ICDCS*, pages 317–326, 2013.

Dataset	Running Time (min)			Memory (MB)		
	TIM$^+$	EaSyIM (l=1)	Gain	TIM$^+$	EaSyIM (l=1)	Gain
DBLP	783.1	2183	0.36x	35234.75	46.5	**758x**
YouTube	NA	5089.5	∞	NA	158.3	∞
socLive	NA	15433.33	**∞**	NA	974.94	∞

Table 2: **Comparing *EaSyIM* with TIM$^+$,** $k = 50$**,** $l = 1$**,** $\epsilon = 0.1$**.**

[2] l is the maximum path length considered for score assignment, where $l \leq \mathcal{D}$. $\mathcal{D}$ is the diameter of the graph.

[3] These instances were run in parallel on 20-cores for *OSIM* and *EaSyIM*. However, for a fair comparison with other techniques we report the total time taken.

[4] Considering the ease of availability of multiple cores in a single machine when compared to large amount of RAM.