IRIE: Scalable and Robust Influence Maximization in Social Networks

Kyomin Jung
KAIST
kyomin@kaist.edu

Wooram Heo
KAIST
modesty83@kaist.ac.kr

Wei Chen
Microsoft Research
weic@microsoft.com

Finding influential users in a social network is essential for viral marketing and social media marketing. Influence maximization problem is defined as finding a node set S of given size K in a social network to maximize their *influence spread* — the expected total number of activated nodes under a certain diffusion process initiated from the set S. In this work, 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 [7, 8] and its extension IC-N [1] 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 [2] for large networks with tens of millions of nodes and edges, while using only a fraction of memory.

In the IC model, each activated node has a single chance to activate each of its outgoing neighbor with a probability assigned to the edge. The IC model can be identified with the Susceptible/Infective/Recovered (SIR) model for the epidemic spreading [10]. Kempe et al. [7] showed that finding optimum solution for the influence maximization under the IC model is NP-hard, and proposed a Greedy algorithm that obtains $(1-1/e)$-approximation for the problem. A number of follow-up works tackle the problem by designing more efficient and scalable optimizations and heuristics [8, 9, 4, 3, 2, 4, 5, 11]. Among them PMIA [2] has stood out as the most efficient heuristic so far.

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. These are especially slow for the first round where the influence spread of every node needs to be estimated. Instead of estimating influence spread for each node at each round, we devise a global influence ranking method, Influence Rank(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 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 contrast, IRIE mainly uses global iterative computations without storing extra data structures, and thus the memory overhead is small.

Let $\sigma(S)$ be the expected total number of activated nodes given a seed set S. When $S = \emptyset$, our algorithm computes estimate $r(u)$ of influence $\sigma(\{u\})$ of a node u by the following equation.

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

where N^{out} is a set of our-neighbor of u, P_{uv} is the probability that u activates its out-neighbor v, and $\alpha \in (0,1]$ is a damping factor. We first prove that $r(u)$ with $\alpha = 1$ is very close to $\sigma(\{u\})$ in any tree graph, and show that $r(u)$ is a good estimate of $\sigma(\{u\})$ in any graph. After selecting some seed node, we consider the influence from the selected seed node. Let $AP_S(u)$ be the probability that a node u is activated when the diffusion process begins from the seed set S. To estimate $AP_S(u)$, we adopt a local tree-based approximation to influence of each seed node [2]. Then, IRIE computes estimate $r(u)$ of marginal influence $\sigma(S \cup \{u\}) - \sigma(S)$ of a node u by the following equation.

$$r(u) = (1 - AP_S(u)) \cdot \left(1 + \alpha \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right) \right). \tag{2}$$

The factor $(1 - AP_S(u))$ indicates the probability that a node u is not activated by the seed set S. Note that (1) and (2) are exactly same when $S = \emptyset$. We compute iterative computations of (2) up to t times and obtain $r^{(t)}(u)$, which computes the estimate of marginal influence of u within distance t from u. Through extensive experiments, we observe that $t = 5$ is good enough for most of the networks.

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 [5], 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 [1]. Our simulation results again show that IRIE has comparable influence coverage while scales much better than the MIA-N heuristic proposed in [1].

Further details of our work including the details of algorithm description and experimental results are presented in our technical report [6].

References

[1] W. Chen, A. Collins, R. Cummings, Te Ke, Zhenming Liu, David Rincon, Xiaorui Sun, Yajun Wang, Wei Wei, and Wifei Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. In *SDM*, 2011.

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

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

[4] 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.

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

[6] K. Jung, W. Heo, and W. Chen. Irie: Scalable and robust influence maximization in social networks. In *arXiv:1111.4795v3*.

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

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

[9] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. In *KDD*, pages 199–208, 2007.

[10] M. Newman. The structure and function of complex network. *SIAM Review*, 45:167256, 2003.

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