Influence in Social Networks

The social networks have grown at an unprecedented level in the last decade where millions of people interact at any given time. Social network has become so pervasive in a way that they could influence the behaviour of individuals and groups that it has led to much research in many areas, particularly information diffusion. It refers to spread of information, through "word of mouth" propagation among friends in social network. Information diffusion has attracted extensive research from many areas including physics [1], economics [2], computer science [3], [4], sociology [5], epidemiology [17] and health [15]. Many applications of where information spread is studies include viral marketing [5], network monitoring [6], rumour control [7], influential spreader in social network [14], social recommendation [8] and causal analysis [16].

A main algorithmic problem in information diffusion is influence maximization (IM). Main goal of IM is to select a set of k users in online social network, (seed set) with the maximum influence spread (i.e. is the maximum no of individuals influenced when the seed set is maximized). The problem was first modelled as probalistic model of interaction with heuristics [9] for choosing the best seeds and then as a discrete optimization problem[3],[4]. Despite many applications of IM it faces many challenges. First is modelling the information diffusion in social network that would heavily affect the influence of any seed set in IM. Second IM problem is computationally intractable. It has been shown that obtaining an optimal solution is NP-Hard [3]. Although a greedy solution has been proposed that achieves 1-1/e approximation [3]. Third, due to random nature of social networks, even evaluation of information spread of an individual seed is computationally complex. Fourth, identifying the influence can be challenging as Individuals tend to engage in similar activities as their peers, so it is often impossible to determine from observational data whether a correlation between two individuals' behaviours exists because they are similar or because one person's behaviour has influenced the other. [16]

With various research in optimization framework of IM various solutions have been proposed, that reduces the run time of the original greedy algorithm [10-12]. Also there has been research to improve that have included network assortavity that have improved that has led to increase in the influence propagation as compared to baseline models. [13]. There have been various models proposed in the literature to model the influence maximization problem, including linear threshold model (LT), independent cascade model (IC), triggering model (TR), time aware diffusion models and non-progressive diffusion models [24], with the most popular being LT and IC.

Threshold model assume that there exists a threshold value for node or people at which they adopt the product or behaviour. In the simplest of these models, the LT model each node v has a latent threshold θ(v) and for every neighbour u∈ N(v) (where N(v) is set of neighbours in the graph) (u,v) has a non-negative weight $w_{(uv)}$ such that $\sum u\, N \in (\,)\, v\, wuv \leq 1$.. The process unfolds in discrete time steps. At time t, an inactive node v becomes active if $\sum_{u\, A \in (\,v)} w_{(uv)} \geq \theta v$ where A(v) is the set of active neighbours of v up to time-step t- 1. Also, it is a type of progressive algorithm so once a node is activated it would remain active. The diffusion process terminates when there are no more nodes remain to activate. Edge weights assigned to threshold models are either assigned with a uniform probability from the set{0.1,0.01,0.001}, or assigning each edge with a constant weight ($1/d^{in}$) where $d^{in}$ is the indegree of the node.

Next famous model is independent cascade model, it considers a user v is activated by each of its incoming neighbours independently by

introducing an influence probability $p_{u,v}$ to each edge e = (u,v). Like LT it is also a discrete optimization model and the process unfolds in discrete time steps. Each active user u in time step t will activate each of its outgoing neighbour with probability $p_{u,v}$.

Each node is given a single chance to activate its outgoing neighbours. After that u stays active and stops the activation. The diffusion instance stops when no more nodes can be activated. $p_{uv}$ parameter in the IC model is assumed to be a constant uniform edge propagation probability of 0.01 or 0.1. The influence spread of seed set S under the IC model is the expected number of activated nodes when S is the initial active node set and above stochastic process is applied.

IC and LT are all time unaware models where the diffusion terminates when no more nodes could be activated. However sometimes the propagation campaigns are time critical and requires maximizing the influence spread under a time constraint. To meet such demands, time-aware models are proposed. So the existing models can be classified as 1) discrete time step models where diffusion happens in discrete time steps 2) the continuous models where the influence spread is in continuous time. The discrete-time models [19], [20], [21] extend IC by modelling the diffusion process from one node to another as a discrete random variable over different time steps. Nevertheless, these models are essentially like IC and LT, as the diffusion only happens in discrete steps. The continuous version of Continuous-time IC considers the likelihood of pairwise propagation between nodes is a continuous distribution of time. Specifically, given the activation time $t_u$ of a node u, the conditional likelihood of u activating its neighbour v at any time $t_v > t_u$ is defined as $p(t_u|t_v; \alpha_{uv})$where $\alpha_{u,v}$ is the parameter of a time-aware influence distribution to determine the influence strength from u to v. Given a predefined stopping time T (T > 0 ), each diffusion process stops when no more node is activated before T.

In opposition to progressive models there also exists non-progressive diffusion models. Typically, non-progressive models are SIR/SIS [22]. There also exists other model like Triggering model [24] which is a neighbouring node is activated based on the set of Triggering nodes $T_v$. IC and LT are all special cases of triggering model (TR).

The process of influence maximization is NP-hard under the LT, IC and TR models. Furthermore, computing the influence σ(S) of a seed set is #P-hard under the LT and IC model. [24]. So there exists greedy framework which guarantees (with approximation ration: 1 – 1/e) termination provided σ(.) is a non-negative monotone and submodular function. The algorithm is initialized with an empty seed set S, and it iteratively selects a node u into S if u provides the maximum marginal gain to the influence function σ(S) wrt. S.

Although the afore mentioned greedy framework has a good approximation ratio of (1-1/e), IM is still very challenging to solve, because evaluating σ(S) is a #P-hard problem. So various approaches have been suggested to overcome that obstacle and they are categorized based on how they overcome this P hardness. These are simulation-based approach, proxy based and sketch based. [24].

Simulation based approach utilize monte-Carlo simulation for evaluating influence spread. It starts from seed set S and traverses G by removing each edge e = (u, v) with a probability $(1 - p_{uv})$ until no user can be reached, resulting in a sample instance. In such a way, we can generate multiple sample instances, and the influence spread σ(S) can be estimated from the sample instances. Although this approach preserves the model generality but it has significant computational overheads.

Proxy based approach believes that the complex influence model can be effectively reduced to proxy models, e.g., PageRank or the

shortest path in practice. Although being efficient it can be unstable under certain condition. [23]. Goal of goal of sketch based approach is to devise theoretically efficient solutions (instead of being only practical efficient) that also preserves a constant approximation ratio, and thus overcome the drawbacks of the above two categories of approaches. For example, the expected time complexity to get a solution in this category [15] is near linear to the size of the input graph with a constant approximation ratio[12].

References:


[1]: Kitsak, M. et al. Identifcation of infuential spreaders in complex networks. Nat. Phys. 6, 888–893 (2010).

[2]: Banerjee, A., Chandrasekhar, A., Dufo, E. & Jackson, M. Te difusion of microfinance. Science 341, 1236498(2013)

[3] Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of infuence through a social network. In Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 137–146 (2003)

[4] Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of infuence through a social network. Teory Comput. 11, 105–147 (2015)

[5] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, “Cost-effective outbreak detection in networks,” in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2007, pp. 420–429

[6] C. Budak, D. Agrawal, and A. El Abbadi, “Limiting the spread of misinformation in social networks,” in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 665–674

[7] X. He, G. Song,W. Chen, and Q. Jiang, “Influence blocking maximization in social networks under the competitive linear threshold, model,” in Proc. SIAM Int. Conf Data Mining, 2012, pp. 463–474

[8] M. Ye, X. Liu, and W.-C. Lee, “Exploring social influence for recommendation: A generative model approach,” in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 671–680

[9] Domingos, P. & Richardson, M. Mining the network value of customers. InProc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 57–66 (2001).

[10] 19. Leskovec, J. et al. Cost-efective outbreak detection in networks. In Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 420–429 (2007).

[11]. Goyal, A., Lu, W. & Laksmanan, L. Celf++: optimizing the greedy algorithm for infuence maximization in social networks. In Proc. 20th International Conference Companion on World Wide Web 47–48 (2011).

[12]. Tang, Y., Xiaokui, X. & Shi, Y. Infuence maximization: near-optimal time complexity meets practical efciency. In Proc. 2014 ACM SIGMOD International Conference on Management of Data 75–86 (2014).

[13] Aral, Sinan, and Paramveer S. Dhillon. "Social influence maximization under empirical influence models." *Nature Human Behaviour* (2018)

[14] Ma, Ling-ling, Chuang Ma, Hai-Feng Zhang, and Bing-Hong Wang. "Identifying influential spreaders in complex networks based on gravity formula." *Physica A: Statistical Mechanics and its Applications* 451 (2016): 205-212.

[15] Fratiglioni, Laura, Hui-Xin Wang, Kjerstin Ericsson, Margaret Maytan, and Bengt Winblad. "Influence of social network on occurrence of dementia: a community-based longitudinal study." *The lancet* 355, no. 9212 (2000): 1315-1319.

[16] Bakshy, Eytan, Itamar Rosenn, Cameron Marlow, and Lada Adamic. "The role of social networks in information diffusion." In *Proceedings of the 21st international conference on World Wide Web*, pp. 519-528. ACM, 2012.

[17] Christakis, Nicholas A., and James H. Fowler. "The spread of obesity in a large social network over 32 years." *New England journal of medicine* 357, no. 4 (2007): 370-379.

[18]Gomez-Rodriguez, M. et al. Infuence estimation and maximization in continuous-

time difusion networks. ACM Trans. Inf. Syst. 34,9 (2016)
[19] B. Liu, G. Cong, D. Xu, and Y. Zeng, “Time constrained influence maximization in social networks,” in Proc. IEEE 12th Int. Conf. Data Mining, 2012, pp. 439–448.
[20] W. Lu, X. Xiao, A. Goyal, K. Huang, and L. V. S. Lakshmanan, “Refutations on ”debunking the myths of influence maximization: An in-depth benchmarking study”,” CoRR, vol. abs/ 1705.05144, 2017, arXiv: https://arxiv.org/abs/1705.05144
[21] J. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-ofmouth,” Marketing Lett., vol. 12, no. 3, pp. 211–223, 2001
[22] W. O. Kermack and A. G. McKendrick , “A contribution to the mathematical theory of epidemics,” in Proc. Roy. Soc. London A:Math., Phys. Eng. Sci., vol. 115, no. 772, pp. 700–721, 1927
[23] X. He and D. Kempe, “Stability of influence maximization,” KDD 2014: pp. 1256-1265
[24] Yuchen Li ; Ju Fan ; Yanhao Wang ; Kian-Lee Tan “Influence maximization on social graphs” A survey in IEEE Vol 30 issue 10.