Influence Maximization on Complex Networks with Intrinsic Nodal Activation

Arun V. Sathanur(✉) and Mahantesh Halappanavar

Pacific Northwest National Laboratory, Richland, WA 99352, USA
arun.sathanur@pnnl.gov

Abstract. In many complex networked systems such as online social networks, at any given time, activity originates at certain nodes and subsequently spreads on the network through influence. Under such scenarios, influencer mining does not involve explicit seeding as in the case of viral marketing. Being an influencer necessitates creating content and disseminating the same to active followers who can then spread the same on the network. In this work, we present a simple probabilistic formulation that models such self-evolving systems where information diffusion occurs primarily because of the intrinsic activity of users and the spread of activity occurs due to influence. We provide an algorithm to mine for the influential seeds in such a scenario by modifying the well-known influence maximization framework with the independent cascade diffusion model. A small example is provided to illustrate how the incorporation of intrinsic and influenced activation mechanisms help us better model the influence dynamics in social networks. Following that, for a larger dataset, we compare the lists of influential users identified by the given formulation with a computationally efficient centrality metric derived from a linear probabilistic model that incorporates self activation.

Keywords: Complex networks · Influence maximization · Social influence · Self-activation · Centrality · Spectral methods

1 Introduction and Related Work

Diffusion of information in complex networks has been the subject of intense scrutiny for researchers and practitioners in many fields. A particular problem that has captured the attention is the identification of central or influential nodes on the network. One rigorous approach to finding influential users with motivations originating in viral marketing is the approach based on influence maximization. We can define the influence maximization problem as follows. Consider a directed graph $G = (V, E)$ that abstracts a complex network, where V is the set of vertices $V = \{v_1, v_2, v_3 \dots\}$ and E is the set of directed edges $\{(v_u, v_w) | v_u, v_w \in V\}$. Further, the vertices are labeled as either *Passive* or *Active* denoting the state of the vertex and a necessary but not sufficient condition for an active vertex v_u to activate a passive vertex v_w is that $(v_u, v_w) \in E$. The other conditions come from the nature of the local diffusion model that is also provided.

E. Spiro and Y.-Y. Ahn (Eds.): SocInfo 2016, Part II, LNCS 10047, pp. 133–141, 2016.
DOI: 10.1007/978-3-319-47874-6_10

Given that there is budget to activate k vertices, the influence maximization problem aims to find that particular set of k *seed vertices* called the *seed set* S, that when activated results in maximal activations on the network amongst all possible such sets of k vertices.

Starting with the landmark paper by Kempe, Kleinberg and Tardos [7], several works have explored newer diffusion models and variations to the ones studied in the work by Kempe *et al.*, namely the independent cascade model and the linear threshold model. These models explicitly address one or more *sociological* aspects of influence. Li *et al.* in [11] consider influence dynamics and influence maximization under a general voter model with positive and negative edges. In a follow-up work Kempe *et al.* [8] discuss a diverse set of models including the so called decreasing cascade model where attempts by multiple neighbors to activate a node results in decreasing probabilities for activation, as the size of the set of neighbors trying to activate the node increases. The authors in ref. [14] propose a general diffusion model that takes into account different granularities of influence, namely pair-wise, local neighborhood etc. The authors in [2], consider influence maximization under the scenario where negative opinions may emerge and propagate. In [5], the authors consider the problem of identifying the individuals whose strong positive opinion about a product will maximize the overall positive opinion about the product. In the process, the authors leverage the social influence model proposed by Friedkin and Johnson [4].

Next we consider the models that address two different types of activation namely intrinsic and influenced. For example, in an online social network (OSN) these can refer to users posting content on their own and users retweeting or liking the posts respectively. Myers, Zhu and Leskovec investigate the diffusion of information, with origins external to that of a social network, through the internal social influence mechanism [12]. In a recent work [3] the authors recognize that the events on social media can be categorized as exogenous and endogenous and model the overall diffusion through a multivariate Hawke's process. While being similar in spirit to these works, our work is more geared towards mining influential nodes in scenarios with intrinsic and influenced activation.

We make the following contributions in this work. Our approach provides for probabilistically modeling the intrinsic and extrinsic activation mechanisms. We then examine these mechanisms in the context of influencer mining from *two different perspectives*, namely the well-known combinatorial influence maximization perspective and a generalized PageRank-type centrality perspective. Carefully chosen experiments on real-world-like and real-world graphs are used to illustrate the two perspectives.

2 Modified Influence Maximization Approach

Considering nodal activation to originate from two specific mechanisms, namely, *Intrinsic* and *Influenced*, allows us to effectively model the so-called *self-evolving* systems such as OSNs that are comprised of content creators (higher probability of getting activated intrinsically) and content consumers (activated via social

influence and spreading the content). Recognizing that most of the users are in some sense content creators and content consumers at the same time, we introduce a real-valued parameter $\alpha \in [0, 1]$ that models the probability of self activation. The total probability for activation for a given node (user) i is a weighted sum of the probability for activation from the two different mechanisms. The parameters $\alpha(i)$ and $\beta(i)$ denote the probability of activation intrinsically and through influence respectively and we have $\alpha(i) \geq 0$, $\beta(i) \geq 0$. The influenced part of the probability for activation is then expressed as a weighted sum of the activation probabilities of the connected neighbors of the user under scrutiny. The w_{ij}s denote the probability of user j activating user i, given that user j is activated by either of the above means. These mechanisms and the associated coefficients are described in Fig. 1. The above probabilistic formulation is similar to the Friedkin-Johnson social influence model for opinion change [4] where the authors recognize that the dynamics of opinion change are governed by two mechanisms - the intrinsic opinion and the influenced opinion.

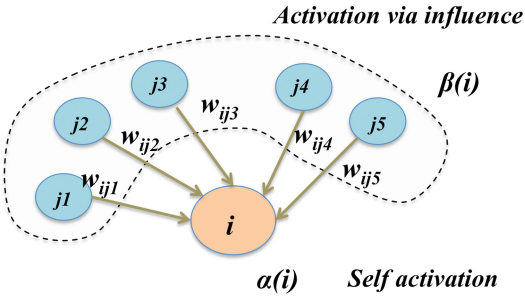

Fig. 1. A concise representation of the self and influenced mechanisms of activation of a node i

Note that all the model parameters discussed above can be efficiently determined as maximum likelihood estimates by observing the activity on the desired portion of the network. For example the proportion of tweets by a user i that are intrinsic in nature can quantify $\alpha(i)$ while a particular weight w_{ij} can be determined by the proportion of user i's retweets (or influenced activity) having their origin in the activity of user j that user i follows. While these *local influence models* can be determined in alternate ways, our focus is to find the overall influencers once these model parameters are estimated.

We propose a simple modification to the classic influence maximization framework using the greedy hill-climbing optimizer [7] working with the independent cascade (IC) model, to incorporate the self-activation mechanism. Let us assume that we are seeking k influential nodes out of a total of N nodes on the network. Let S^p be the set of influential nodes at step $p \leq k$. The greedy hill-climbing optimizer expands the set to size $(p + 1)$ by polling each of the vertices not in S^p and augmenting those vertices, one at a time to form the set

$S^p \cup v$ and looking for the best marginal gain in terms of the activations. At each such step p, instead of setting each of the nodes in $S^p \cup v$ to be activated and then computing the activations according to the IC model, we probabilistically activate each node in $S^p \cup v$ with a probability given by the corresponding α values to simulate the intrinsic activation process. This modification is depicted in Algorithm 1. Given the probabilistic nature of the algorithms, the overall activation numbers are obtained by running the diffusion model in a Monte Carlo fashion by invoking n independent trials involving randomized graphs with corresponding edge weights. We further accelerate the process of finding the influential nodes by parallelizing the Monte Carlo runs by leveraging *multi-threaded platforms*.

Algorithm 1. Selects a set of k influential nodes that cause maximal activations on a network, following the independent cascade(IC) model with self-activation (SA). The inputs are a directed graph ($G = (V, E)$), set of edge weights ($P = \{p_{uv} : (uv) \in E\}$), vector of alpha values ($\alpha = \{\alpha_v : v \in V\}$), number of samples ($n$), and number of seeds to be identified (k).

procedure IC-SA(G, P, α, k, n)
  Generate n random numbers r^1_{uv} ...r^n_{uv} for each edge in E and generate a set SG containing n subgraphs such that in subgraph i, $p_{uv} \geq r^i_{uv}$
  $S \leftarrow \emptyset$ ▷ Set of influential nodes to be mined
  while $|S| < k$ **do**
    $v_{best} \leftarrow \emptyset$, $a_{best} \leftarrow 0$
    for each node v in $V \setminus S$ **do**
      $a \leftarrow 0$
      for each $G_i \in SG$ in `parallel` **do**
        $\hat{S} \leftarrow$ active nodes in $S \cup \{v\}$ based on α
        Compute number of nodes, $\hat{a}$, in $V \setminus \hat{S}$ that are reachable from the $\hat{S}$
        $a \leftarrow a + \hat{a}$ ▷ Synchronized update
      if $a \geq a_{best}$ **then**
        $v_{best} \leftarrow v$
        $a_{best} \leftarrow a$
    if $v_{best} \neq \emptyset$ **then**
      $S \leftarrow S \cup \{v_{best}\}$
  return S

We also adopt the weighted-cascade method for normalizing the edge probabilities [7]. Thus if $\boldsymbol{W}$ denotes the sparse weight matrix that characterizes the IC edge probabilities, we require that $\boldsymbol{W}$ be row-stochastic. That is $\sum_{j,(j,i)\in\mathcal{E}} w_{ij} = 1$. Further by assuming that the nodes are not *lazy* and are activated by either of the two mechanisms that we outline, we set $\beta(i) = (1 - \alpha(i))$. This will render the overall IC probability between nodes j and i to be $(1 - \alpha(i))w_{ij}$. The assumption that $(\alpha(i) + \beta(i)) = 1$ is being relaxed in the ongoing work where we allow $(\alpha(i) + \beta(i)) \leq 1$ thereby modeling the slack with a *lazinesss* factor.

3 Experiments

3.1 Small Organization Tree

We first consider a small directed and weighted network with 23 nodes, organized in a tree-like fashion. The graph is depicted on the left side of Fig. 2. In this experiment, we consider the tree-like network to depict a small organization with a Director (Node D), two Managers (M1 and M2) and twenty Employess (E1-E20), with 10 employees each working under the two managers. We set $\alpha^0(D) = 0.95$ signifying that the Director almost exclusively acts intrinsically. We also set $\alpha^0(M1) = \alpha^0(M2) = 0.25$. All the employees have an α of 0.25 as well. As for the weights, the edges ending at node D receive weights of 0.5 each (when the Director chooses to be influenced, the director gets influenced equally by the two managers). As for the managers, they have a weight of 0.5 each on the edges that are incoming from D and the remaining 0.5 is split equally among the edges originating at the 10 employees each. The employees carry a weight of 1.0 on the edges originating from their managers. We then perturb this baseline case to mimic a situation where the director starts becoming more susceptible to influence while the manager $M1$ starts becoming inflexible. This is done by setting $\alpha(D) = \alpha^0(D) - \delta$ and $\alpha(M1) = \alpha^0(M1) + \delta$. We then sweep δ from 0.05 to 0.45. The results are shown in the right panel of Fig. 2 where we can see that D starts out as the most influential node as expected, but then M1 becomes more influential than D at a certain value of δ and will eventually have reach over all the employees on the network. Note that the activation numbers plotted on the y-axis are the cumulative activations over all the $n = 3200$ Monte Carlo samples. This simple experiment shows that influence on social networks is sensitive to the extent of intrinsic activation and clearly such scenarios cannot be easily captured by the traditional influence maximization framework.

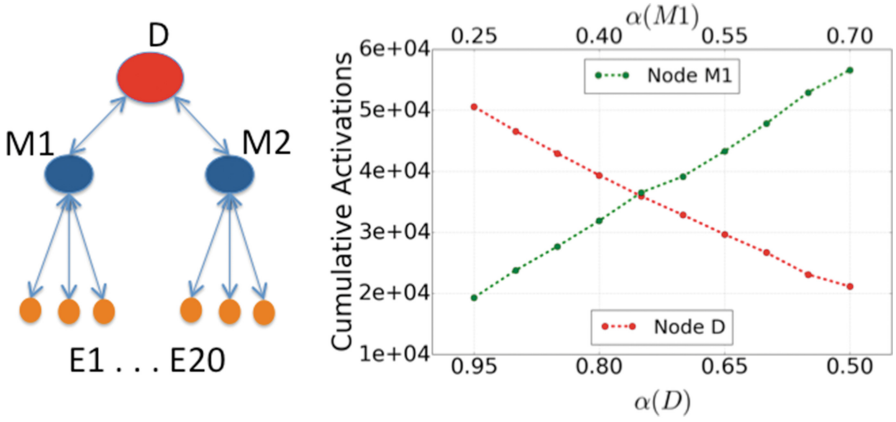

Fig. 2. The small organizational tree network (left) and the behavior of the influence functions with the various α values.

3.2 Larger Graphs and the Influence Function

Our dataset of larger graphs consists of

- LFR-1000 graph with 1000 vertices and 11433 edges is a synthetic network that follows the generative LFR model that mimics real-worlsd graphs [10] .
- The PBlogs graph [1] that represents a real-world blogs network and has 1095 vertices and 12597 edges.

The details of these graphs are discussed in [6]. Visualizations of the two larger graphs are shown in the inset in Fig. 3.

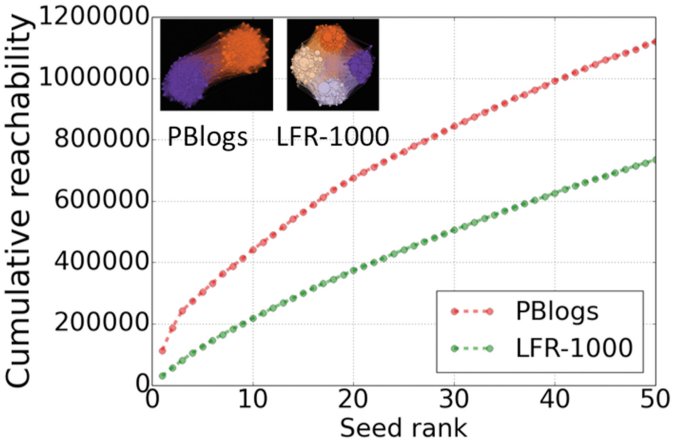

Fig. 3. Submodular nature of the influence function under self-activation. *Inset: The PBlogs (left) and LFR-1000 (right) networks visualized in Gephi*

For the classic influence maximization problem with the independent-cascade model, the greedy hill-climbing optimizer is shown to be optimal in the sense that it provides a $(1 - \frac{1}{e} - \epsilon)$ approximation guarantee because the expected influence spread is a sub-modular function. For the case of intrinsic nodal activation, we have not been able to prove the sub-modularity of the influence function. While leaving the formal proof as an open problem, we conjecture that the sub-modularity holds because for each seed l selected, we can introduce an edge pointing from a dummy node to the seed l with an activation probability equal to $\alpha(l)$ and seed the dummy node. This process does not interfere with the normal operation of the network except activating the seed l with a probability $\alpha(l)$, exactly as required and therefore preserves the sub-modularity of the influence function. When we applied the modified influence maximization approach given by Algorithm 1 to the LFR-1000 and the PBlogs graphs and requested for 50 seeds, we observed from Fig. 3 that the cumulative influence spread (total number of activations from all the samples considered) showed a sub-modular character as evidenced by the diminishing gains [9] in the total number of activations for each seed added.

4 Comparison with an Equivalent PageRank-Type Influence Measure Based on a Linear Model

We examine the influencer mining on networks with intrinsic and influenced nodal activations from a slightly different perspective. By collecting the various probabilities together and recognizing the recursive nature of influence spread on a social network, we arrive at a generalized, computationally efficient, PageRank type spectral influence measure that incorporates the two activation mechanisms [13]. As demonstrated in [13], when considering activity on an OSN, this approach is a better measure of influence spread than a purely topological metric such as PageRank.

For a given node i, from Fig. 1, the total probability of activation $p_A^T(i)$ can be written as

$$p_A^T(i) = \alpha(i) + \beta(i) \left(1 - \prod_{j,(j,i)\in\mathcal{E}} \left(1 - w_{ij} p_A^T(j)\right) \right) \tag{1}$$

By retaining the leading-order terms, we get a linearized version of Eq. 1 as

$$p_A^T(i) = \alpha(i) + \beta(i) \sum_{j,(j,i)\in\mathcal{E}} w_{ij} p_A^T(j) \tag{2}$$

$p_A^T(i)$ denotes the total probability of activation for node i (intrinsic and influenced). The parameters $\alpha(i)$ and $\beta(i)$ denote the weights for activation intrinsically and through influence respectively as before and we set $\beta(i) = (1 - \alpha(i))$ as before. We now extend Eq. 2 to the entire network with N nodes to obtain a matrix-vector equation.

$$\boldsymbol{p_A^T} = \boldsymbol{\alpha 1} + ((\boldsymbol{I} - \boldsymbol{\alpha})\, \boldsymbol{W})\, \boldsymbol{p_A^T} \tag{3}$$

Here $\boldsymbol{p_A^T}$ is a vector of size $N \times 1$ denoting respectively the total probability of activation for all the nodes on the network. $\boldsymbol{I}$ denotes the identity matrix of size $N \times N$. $\boldsymbol{\alpha}$ denotes the diagonal matrix with entries corresponding to the intrinsic activation probability for all the nodes on the network. $\boldsymbol{W}$ denotes the sparse, stochastic weight matrix with entires given by the w_{ij}s discussed earlier. $\mathbf{1}$ is the all-ones vector of size $N \times 1$

We can then express the total activation probabilities as

$$\boldsymbol{p_A^T} = \mathbf{1}^T \boldsymbol{G}; G = (\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{\alpha})\boldsymbol{W}))^{-1}\, \boldsymbol{\alpha} \tag{4}$$

Here $\mathbf{1}^T$ is utilized to give us the column-sum of $\boldsymbol{G}$. We also note that because the matrix $\boldsymbol{W}$ is a row-stochastic matrix, the matrix $\boldsymbol{G}$ is also row-stochastic. The quantity $C_A(i) = \left(\sum_{j=1}^{N} G_{ji}\right)$ which corresponds to the sum of the entries in column i of $\boldsymbol{G}$, represents the expected number of hosts activated by node i and is a measure of influence. We term this *amplification factor* as *activation centrality* and it can be directly computed as a linear solve.

Table 1. Correlations, two ways, between the proposed approaches for the two inputs PBlogs and LFR1000 for different sizes of seed sets (10, 20, 30, and 50). Closer the metric to one the better.

Correlation type	Input	$k = 10$	$k = 20$	$k = 30$	$k = 50$
Jaccard	PBlogs	0.538	0.818	0.875	0.818
RBO	PBlogs	0.817	0.846	0.851	0.868
Jaccard	LFR1000	0.818	0.905	0.765	0.818
RBO	LFR1000	0.979	0.963	0.947	0.937

In our experiments, the $\boldsymbol{\alpha}$ and $\boldsymbol{W}$ entries were randomized with entries drawn from the uniform distribution over $[0, 1]$ and $\boldsymbol{W}$ was converted to a row-stochastic matrix. We then compare the sets of top-k influencers identified by both the methods on two larger graphs in our dataset. The comparison was carried out with respect to two measures - Jaccard similarity and the rank-biased overlap (RBO) that also considers ordering with higher weights given to matches that happen at the top [15]. These results are presented in Table 1 where we see excellent agreement between the sets of influential nodes obtained by both the methods. Thus the activation centrality metric which includes the *intrinsic activation* mechanism represents a computationally more viable alternative to the full-scale influence maximization framework, retaining the essence of the model and being amenable to a sparse matrix based linear solve.

5 Conclusions and Ongoing Work

In this short paper we introduced the notion of vertices on a social graph getting activated by two mechanisms, namely, intrinsically and through social influence as commonly observed in online social networks. Utilizing a modified version of the influence maximization framework that combines the self-activation probability with the independent cascade model for influenced activation, we were able to find the influential nodes on such a network. We then introduced a spectral centrality measure of influence that takes into account, the intrinsic and influenced activation mechanisms and demonstrated that the sets of influential users identified by the two mechanisms agree very well. Building on this preliminary work, ongoing work is exploring multiple facets of this problem including the exploration of how a social network can be successful in the long run by balancing the two modes of activation. We are also extending these methods other complex systems such as for attack modeling in cyber networks.

Acknowledgement. This research was supported by the High Performance Data Analytics program at the Pacific Northwest National Laboratory (PNNL). PNNL is a multi- program national laboratory operated by Battelle Memorial Institute for the US Department of Energy under DE-AC06-76RLO1830.

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 us election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 36–43. ACM (2005)
2. Chen, W., Collins, A., Cummings, R., Ke, T., Liu, Z., Rincon, D., Sun, X., Wang, Y., Wei, W., Yuan, Y.: Influence maximization in social networks when negative opinions may emerge and propagate. In: SIAM Data Mining, pp. 379–390 (2011)
3. Farajtabar, M., Du, N., Gomez-Rodriguez, M., Valera, I., Zha, H., Song, L.: Shaping social activity by incentivizing users. In: Advances in Neural Information Processing Systems, pp. 2474–2482 (2014)
4. Friedkin, N.E., Johnsen, E.C.: Social influence networks and opinion change. Ad. Group Proces. **16**(1), 1–29 (1999)
5. Gionis, A., Terzi, E., Tsaparas, P.: Opinion maximization in social networks. In: SIAM Data Mining Conference, pp. 387–395. SIAM (2013)
6. Halappanavar, M., Sathanur, A., Nandi, A.: Accelerating the mining of influential nodes in complex networks through community detection. In: Proceedings of the 13th ACM International Conference on Computing Frontiers, CF 2016, Como, Italy, May 16–18, 2016
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of ACM SIGKDD, pp. 137–146. ACM, New York (2003)
8. Kempe, D., Kleinberg, J., Tardos, É.: Influential nodes in a diffusion model for social networks. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) ICALP 2005. LNCS, vol. 3580, pp. 1127–1138. Springer, Heidelberg (2005). doi:10.1007/11523468_91
9. Krause, A., Golovin, D.: Submodular function maximization. Tractability Pract. Approaches Hard Probl. **3**(19), 8 (2012)
10. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**(1), 016118 (2009)
11. Li, Y., Chen, W., Wang, Y., Zhang, Z.L.: Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 657–666. ACM (2013)
12. Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: ACM SIGKDD, pp. 33–41. ACM (2012)
13. Sathanur, A.V., Jandhyala, V., Xing, C.: Physense: Scalable sociological interaction models for influence estimation on online social networks. In: IEEE International Conference on Intelligence and Security Informatics, pp. 358–363. IEEE (2013)
14. Srivastava, A., Chelmis, C., Prasanna, V.K.: Influence in social networks: A unified model? In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 451–454. IEEE (2014)
15. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. (TOIS) **28**(4), 20 (2010)