March 1, 2023

To: Professor Jie Lu
Editor in Chief, *Knowledge-Based Systems*

Dear Professor Lu,

We would like to thank you for providing us with the opportunity to revise our paper for the second time (Bozorgi et al, “Community-based Influence Maximization in Social Networks under a Competitive Linear Threshold Model”). As per your suggestion, we have thoroughly revised our paper in response to the detailed comments provided by the anonymous reviewers. Below is given descriptions of how we have addressed each of the reviewer comments.

Sincerely,
Arastoo Bozorgi, Saeed Samet, Johan Kwisthout, Todd Wareham

Reviewer #1

Comment #1: This is the second round of review for manuscript entitled ”Community-based Influence Maximization in Social Networks under a Competitive Linear Threshold Model”. Through its revision, the manuscript has been substantially improved and the authors have addressed some of my comments. However, the main issues raised in connection with the experimental part of the paper are not sufficiently addressed. The comparisons and the analysis of results is not yet convincing. For instance, the method used for comparative purposes come from only Monte-Carlo simulation. Nevertheless, there are a lot of methods for influence maximization published in recent years, including 2016. The proposed approach needs to be validated in comparisons with such recent works.

Answer: We thank the reviewer for suggesting this revision. We added four more algorithms (INCIM [1], IPA [2], LDAG [3] and HighDegree [4]) to our experiments and compared their results in subsection **Seed selection** of Section 4.2. Both the INCIM and IPA algorithms use the idea of communities to find influential nodes and, like LDAG, have reasonable running times and find nodes with good quality. HighDegree is also a well-known algorithm which is compared with most influence maximization approaches. So, in this version of the paper, we have 6 different set of experiments to simulate the competitive condition from the follower’s perspective (Figure 6). In each experiment, we chose the seed nodes for the first competitor by running one the mentioned algorithms and then we ran our algorithm to choose the minimum number of seed nodes for the second competitor to defeat the first competitor. The analysis are presented in two paragraphs in subsection **Seed selection** of Section 4.2 as follows:

> "To simulate the competitive condition from the follower's perspective, we chose some seed nodes randomly and activated them for the first competitor as negative and ran CI2 to select the minimum number of nodes with higher influence spread for the second competitor. The nodes selected for the second competitor should be different from the ones selected for the first competitor. We also did the same process by running the greedy approximation algorithm, INCIM [1], IPA [2], LDAG [3] and HighDegree [4] algorithms with different values for k as their budgets. The generated seed sets are of size 5, 10, 20, 30, 40 and 50. The minimum number of nodes selected by CI2 to defeat the first competitor are shown in Figure 6.
>
> As we can see in Figure 6, the minimum number of nodes which is required to be selected by the second competitor to achieve higher influence spread depends deeply on how the seed nodes are selected by the first competitor. In Figure 6(a), in which the seed nodes of the first competitor are selected randomly, in each set, fewer number of nodes are required to defeat the first competitor. But when we extract the actual seed nodes by running the mentioned algorithms, in each set of nodes, more nodes need to be selected by the second competitor. For example, in the Slashdot dataset in Figure 6(b), 72 seed nodes need to be selected to achieve higher influence spread than the spread achieved by an actual seed set of size 50, while there only 13 nodes need to be selected to achieve higher influence spread when the nodes in set of size 50 are selected randomly. Also, the algorithm which is used to extract the actual seed nodes for the first competitor, affect the number of seed nodes required to be selected by the second competitor as different algorithms achieve different level of qualities in their seed node extraction. A good comparison of quality of seed nodes extracted by the mentioned algorithms is done in [1]. In Figures 6(c-f), the number of seed nodes that need to be selected to defeat the first competitor are 69, 67, 64 and 61 if the seed sets of size 50 are extracted by the INCIM, IPA, LDAG and HighDegree algorithms respectively from the Slashdot dataset. These Figures tell us that as seed nodes are selected with higher quality for the first competitor, more seed nodes other than the selected ones need to be selected by the second competitor."

Reviewer #3

Comment #1: In social media, the determination of user's activeness only relies on network structure is unreasonable, without considering other social relations.

Answer: We agree with the reviewer that an ideal model of influence propagation should base user activity on more than static network structure. However, it is difficult to know exactly how to incorporate other social relations fully into such a model. We have instead chosen to build on the popular LT model in an incremental (but, we hope we have shown, useful) fashion by incorporating the ability for users to deliberate for a fixed amount of time over which influence to adopt.

Comment #2: Literature [22, 23] show that the customer's choice among competing products relies more on recent than old information. The submission tried to apply the information of d time steps. The authors need to explain how to reasonable use this information.

Answer: We agree with the reviewer to clarify the effect of d recent time steps. So, we added an example to subsection 3.2 of the paper and compared our approach which consider the information from d previous time steps with two WPCLT and K-LT approaches and explain each approach according to the condition which is depicted in Figure 1. The added paragraph is as follows:

> "As an example, imagine the graph in Figure 1 is part of a social graph in which, nodes w_1 and w_3 have been activated before as *active*$^+$ and *active*$^-$ respectively and nodes w_2 and w_4 haven't been activated yet. If $\theta_v = 0.25$, in time step 1, node v gets influenced in both the WPCLT and K-LT models, as the total incoming influence weight is bigger than its threshold value, and then node v would be activated as *active*$^+$ as $p_{w_1,v} > p_{w_3,v}$. Now imagine that nodes w_2 and w_4 are activated as *active*$^-$ by other nodes of the graph in time step 2. This means that in both the WPCLT and K-LT models, node v is in state *active*$^+$ in time step 2, while the majority of its neighbors have been activated as *active*$^-$. But in DCM, the state of node v changes from *inactive* to *thinking* in time step 1 and its state remains stable for d time steps so that it can consider different influence spreads, after which it decides to be activated by the influence spread which is accepted by the majority of its neighbors. This causes the state of node v to be changed from *thinking* to *active*$^-$ after d time steps in the DCM model. Therefore, it is reasonable to give the ability to nodes to think about the incoming influence spread."

Comment #3: Literature et al. [10] solve the competitive influence maximization problem from the host's perspective, and Literatures [6, 7] look at this problem from the follower's perspective. The submission also studied competitive influence maximization from the follower's perspective. The innovation is not really obvious.

Answer: Thank you for this comment. The main innovation of this paper which studies competitive influence maximization from the follower's perspective is the ability of nodes to think about the incoming influence in a competitive version of LT propagation model, which results DCM model. Also, selecting the seed nodes for the second competitor with spending the minimum budget inside the communities is another main contribution of the paper. We mention these innovations in the first and third items describing our paper's contribution as follows:

> "Item1: We propose the DCM propagation model, which gives decision-making power to nodes based on the incoming influence in a competitive version of the LT propagation model.
>
> Item 3: We propose the CI2 algorithm to find the minimum number of the most influential nodes for a competitor C_2. This algorithm uses knowledge of the nodes selected by a competitor C_1 so that C_2 can achieve more influence spread by spending less budget. Computing the spread of seed nodes is done locally inside communities of the input graph, which we show results in a substantial decrease in running time."

Reviewer #4

Comment #1: Although the authors have made significant improvements to the paper, the impact of this work remains somewhat unconvincing. This may be since the algorithms are still not sufficiently detailed. It would be useful to have some context for how the competitive knowledge of knowing the seeds of the competitor is considered? The motivation of the approach could be improved.

Answer: Thanks for reminding us for this insufficiency in our explanations. We changed the second paragraph of subsection **Seed selection** in section 3.3 to explain the seed selection step in more detail. The revised paragraph is as follows:

> "In each community C_i, we locally run the simple greedy algorithm [1] which uses the DCM model as its propagation model to find the most influential node in C_i and store the node ID and its spread value in candidate seed set S'. Note that the node which is selected as a candidate node in this step should be different from the nodes which have been selected for the first competitor. The size of S' is equal to the number of communities and in each step, this set is updated to hold the new candidate

> seeds of each community. Among the candidate seeds, the one which has the maximum marginal gain is selected and added to S_2. S_1 and S_2 are seed sets of the first and second competitors respectively."

Comment #2: A brief description of the LT algorithm in the background would be beneficial so as to demonstrate the differences and contributions of your proposed method.

Answer: Thanks you. We changed the first paragraph of section 2 as following to explain LT model in more details:

> "In [1], Kempe et al. introduced two propagation models to address the influence maximization problem, the Linear Threshold (LT) and Independent Cascade (IC) models. In both models, a threshold value $\theta \in [0, 1]$ is assigned to each node and each node can be active or inactive. Also, each edge from node u to node v has an influence weight $p_{u,v} \in$ (0,1]. At first, all nodes are inactive except the nodes in set S which have been activated before as seed nodes and the propagation process is started from them. In (LT), an inactive node u can be activated in time t if $f_v(S) > \theta_v$, where S stands for v's neighbors which are activated at time $t - 1$ and, as is mentioned in [1], the value of f_v is initialized as
>
> $$f_v(S) = \sum_{u \in S} b_{v,u}$$
>
> $b_{v,u}$ is the weight of edge (v, u). In the LT model, the sum of all edge weights between v and its neighbors should be less than 1 [1].
>
> In IC, the activation process is the same with LT except that in IC, an activated node u has only one chance to activate its inactive neighbor v with probability $p_{u,v}$."

Comment #3: Contribution, point 3 should be rewritten for clarity. It is a rambling sentence and your points get lost. Also, it would be better to quantify the "remarkable" decrease in execution time.

Answer: As suggested, point 3 has been rewritten for clarity (see Comment #3 for Reviewer #3 above for the revised text). We have also eliminated "remarkable" from the described decrease in running time, deciding instead to let the reader judge the extent of the decrease as described on pages 29 and 30 of the revised manuscript.

Comment #4: point 4: It would be better to describe the data sets to make these points more interesting and clear.

Answer: We thank the reviewer for pointing this out. By adding a brief description of the data sets to this point, we can attract the reader. hence, we revised this point as following:

> "We conduct experiments using three real and three synthetic datasets to show that CI2 can find influential nodes efficiently in an acceptable running time. Synthetic datasets are generated with same number of nodes and edges but different community structures in order to track the effect of community structure of networks on our approach."

Comment #5: p 6 - (1) is expressed after emphasizing thinking state for d timestamps, but $p_{u,v}$ does not refer to d, which is confusing.

Answer: You are right about the confusing explanation. We changed the paragraph as follows:

> "In the DCM propagation model, each node can be in one of the following states: *inactive*, *thinking*, *active*$^+$ and *active*$^-$. Suppose there are two competitors who try to advertise for their products over a social network. We denote the first competitor with the $+$ sign and the second competitor with the $-$ sign, and each node v picks a threshold value θ_v uniformly at random from [0,1]. Let S_1 be the seed set selected by the first competitor and S_2 be the seed set selected by the second one. At first all the nodes except the seed set's nodes are *inactive*. The activation process of node v is as follows: at time $t > 1$ if the total incoming influence weight from the in-neighbors of v which are active ($N^{in}_{active}(v)$) reaches the threshold value of v, its state changes to *thinking*, which means the state of node v changes with probability
>
> $$\sum_{u \in N^{in}_{active}(v)} p_{u,v} \geq \theta_v \tag{1}$$
>
> Node v remains in *thinking* state after this state change for d steps. After that, it decides to become *active*$^+$ or *active*$^-$ based on the maximum total incoming influence weight from its in-neighbors. Let A^+_{t+d} be the set of in-neighbor nodes of v with state *active*$^+$ and A^-_{t+d} be the set of in-neighbor nodes of v with state *active*$^-$ at time $t + d$. The state of node v changes from *thinking* to *active*$^+$ or *active*$^-$ as follows:
>
> $$v_{state} = \begin{cases} active^+, & \text{if } \sum_{u \in A^+_{t+d}} p_{u,v} > \sum_{u \in A^-_{t+d}} p_{u,v} \\ active^-, & \text{otherwise} \end{cases} \tag{2}$$
>
> "

Comment #6: opposite of the recency effect": seems opposite is very specific – as in not recency or old. As I understand it simply does not guarantee that it is recent. So, not such a strong statement may be more accurate, such as does not assure recency, which is often significant when making choices.

Answer: We revised the mentioned sentence to address your point as follows:

> "Wei Lu et al [10] noted that in the WPCLT model, when a node is about to activate, the neighbors which have been activated in all previous time steps are considered; this, however, does not assure recency, which is when the customer's choice among competing products relies more on recent than old information [22,23]."

Comment #7: This notion of solvability is essentially that promised by the various flavours of evolutionary computation" ¡- essentially what vairous flavours of evolutionary computation promises.

Answer: We agree that the initial phrasing was awkward. This sentence has been revised as follows: "This notion of solvability is essentially what many types of stochastic heuristics (in particular, those based on evolutionary computation) promise."

Comment #8: Figure 1 - Overview of our community-based ...

This figure is presented as an overview of the proposed algorithm. As such it should be more clear. The figure caption and referring text should be improved. If it is going to be printed in color, then the different colors and line styles should be explained in caption. If it is not going to be published in color, then it is hard to see. There are no weights represented on the graphs nor explanation of the how the seeds are selected. The lower graph is not labeled.

Answer: Thank you for this point about the color problems in different printed versions and the captions. We added some sentences to briefly explain an overview of our algorithm with reference to Figure 2 in the first paragraph of section 3.3. The revised paragraph is as follows:

> "Motivated by the useful characteristics of communities in social networks which we mentioned previously in Section 2, we decided to base our CI2 algorithm for competitive influence improvement on influential nodes in the community structure of the input graph G. An overview of CI2 is shown in Figure 2. At first, the communities of the input graph are extracted (denoted by labels C_1, C_2 and C_3 in Figure 2). Then, inside each community, the most influential node is selected as a seed candidate. Finally, the node which has the maximum influence spread among candidate nodes is selected as a seed node. The

> selected seed node for the second competitor is denoted b y a + sign in Figure 2. The CI2 algorithm is explained in more detail in the following section."

Also, we modified the caption of Figure 2 and changed the line colors to some patterns so that the differences can be seen in the non-color version.

Comment #9: Table 2: Average values and variance would be interesting to know. More information about the communities and the community structure would be interesting.

Answer: We thank the reviewer for this suggestion. We added two more rows to Table 2 including information about average and maximum out-degrees of the biggest community for all the real datasets. You can see Table 2 with the added information on page 18 on the current version of the paper.

Comment #10: Table 3. Since it is a difference it is important to indicate clearly. The graph indicated spread is less in CI2, so CI2 - MC would be negative. Just need to fix the caption to be clear.

Answer: We thank the reviewer for this suggestion. We changed the table caption as follows:

> "Table 3: Differences between the spread values computed by MC and CI2 (in percentage)"

Editorial comments: Thank you for your detailed revision and your valuable editorial suggestions on the paper. We proofread the paper and we applied all the suggested sentences and words to this version of the paper. Also, we rewritten the long sentences for more readability.

References

[1] A. Bozorgi, H. Haghighi, M. S. Zahedi, M. Rezvani, INCIM: A community-based algorithm for influence maximization problem under the linear threshold model, Information Processing & Management (2016).

[2] J. Kim, S.-K. Kim, H. Yu, Scalable and parallelizable processing of influence maximization for large-scale social networks?, in: Data Engineering (ICDE), 2013 IEEE 29th International Conference on, IEEE, 2013, pp. 266–277.

[3] W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in: Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, 2010, pp. 88–97.

[4] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2003, pp. 137–146.