

An Influence Maximization based approach to the Engagement of Silent Users in Online Social Networks

EXTENDED ABSTRACT

Roberto Interdonato^{1,2}, Chiara Pulice³, and Andrea Tagarelli¹

¹ DIMES, University of Calabria, Italy

² Department of Information Technology, Uppsala University, Sweden

³ UMIACS, University of Maryland, USA

Abstract. We discuss a targeted influence maximization (TIM) approach, based on the linear threshold (LT) diffusion model, which is designed to support the engagement of silent members, a.k.a. lurkers, of an online social network. A key aspect of this approach is the objective function of the underlying optimization problem, which is defined in terms of the lurking scores associated with the nodes in the final active set, or delurking capital. The resulting TIM problem shares the computational intractability with classic influence maximization (IM) problems. However, since the proposed objective function is shown to be monotone and submodular under the LT model, a greedy algorithm (with a guaranteed approximation ratio) is provided to compute a k -node set that maximizes the delurking capital in the network, for a given lurking threshold. Results have demonstrated the significance of our approach. A comparative evaluation with state-of-the-art algorithms for non-targeted and targeted LT-based IM has also shown the superiority of our approach in terms of the estimated overall delurking capital.

1 Introduction

Computer science research in online social networks (OSNs) has traditionally focused on users that are actively involved in the OSN life. However, online social environments are often characterized by the presence of a crowd which does not actively contribute, rather it takes a silent role. Such silent members of an OSN are also known as *lurkers*, since they gain benefit from others' information without significantly giving back to the community [8,2,9]. The study of lurking behaviors in OSNs is nonetheless important, since lurkers acquire knowledge from the community, and as such they can be social capital holders. Within this view, a major goal is to *delurk* such users, i.e., to encourage them to more actively be involved in the OSN. This can in principle be accomplished by nurturing and persuading lurkers to be more actively engaged in the community, also thanks the support of other, possibly influential users [9].

In [5] we proposed the first computational framework that addresses the problem of delurking in OSNs. Our key idea was to conceptualize a delurking approach under a graph-based *information diffusion model*. Research on information diffusion in OSNs (e.g., [1,14,4]) is well-established due to a plethora of methods that have been developed in the last years, mostly upon the two seminal models, namely Independent Cascade

(IC) and Linear Threshold (LT) [6]. These assume that an information diffusion process would unfold in a static, directed graph, where each node can be “activated” or not (under a progressive assumption) and each edge is associated with a weight expressing the diffusion probability or influence degree from one node to another. Moreover, in an *influence maximization* (IM) framework (e.g., [6,3,13]) the general objective is to find a set of initially activated users (also called early adopters, or seed users) which can maximize the spread of information through the network.

Our approach refers to a special case of IM, namely *targeted IM*; in particular, a specific instance of targeted IM, in which lurkers are regarded as the target of the diffusion process. Existing *lurker ranking* algorithms [10,11,12] can profitably be exploited to mine lurking behaviors in the network, and hence to associate users with a value (*lurking score*) indicating her/his lurking status. Given a budget k , our goal is to find a set of k nodes that are capable of maximizing the likelihood of “activating” the target lurkers. Intuitively, the activation of a node means that a user is influenced by other users so to “become aware of” or “adopt” a piece of information. Note that this outcome of a diffusion process has a nice analogy in the lurking analysis setting under consideration: the information to spread corresponds to a delurking trigger action, activating nodes regarded as lurkers can be seen as delurking those users, while the lurking score of a user would represent the gain deriving from delurking that user.

A key novelty in the formulated optimization problem is the objective function, which is defined in terms of the cumulative amount of the lurking scores associated with the nodes in the final active set, or *delurking capital*. Our delurking-oriented targeted IM problem shares the computational intractability with classic IM problems. However, since the proposed objective function is shown to be monotone and submodular under the LT model, we provide a greedy algorithm (with typical $1 - 1/e - \epsilon$ approximation ratio), named DEVOTION, which computes a k -node set that maximizes the delurking capital in the network, for a given minimum lurking threshold.

We evaluated DEVOTION over OSN datasets of different characteristics and sizes, assessing its performance in terms of estimation of the delurking capital, execution time, and seed characteristics. We also compared DEVOTION with state-of-the-art IM algorithms, namely SimPath [3] and TIM+ [13], and a query-based targeted IM algorithm, namely KB-TIM [7].

2 Delurking-oriented Targeted Influence Maximization

2.1 Problem statement

Let $\mathcal{G}_0 = \langle \mathcal{V}, \mathcal{E} \rangle$ be the directed graph representing a social network, with set of nodes \mathcal{V} and set of edges \mathcal{E} . We denote with $\mathcal{G} = \mathcal{G}_0(b, \ell) = \langle \mathcal{V}, \mathcal{E}, b, \ell \rangle$ a directed weighted graph representing the information diffusion graph associated with the social network \mathcal{G}_0 , where $b : \mathcal{E} \rightarrow \mathbb{R}^*$ is an edge weighting function, and $\ell : \mathcal{V} \rightarrow \mathbb{R}^*$ is a node weighting function.

We assume that the node and edge weighting functions in the diffusion graph \mathcal{G} are completely specified based upon the availability of a *lurker ranking* solution. This is produced by a method that is designed to identify and characterize lurkers in an OSN,

i.e., a method enabled to assign every user with a score that expresses the status of lurking behavior taken by that user. The intuition is that for any edge (u, v) in \mathcal{G} , the weight $b(u, v)$ will indicate how much node u has contributed to the v 's lurking score calculated by the lurker ranking method, which resembles a measure of "influence" produced by u to v . Also, the node weight $\ell(v)$ will indicate the status of v as lurker, such as the higher the lurker ranking score of v the higher should be $\ell(v)$. We refer the reader to [11,10] for details about the lurker ranking method and to [5] for definitions of node and edge weighting functions.

We denote with $LS \in [0, 1]$ a threshold value that indicates the *minimum lurking score* that a node in the network is required to have in order to be regarded as a target node. Moreover, given any seed set $S \subseteq \mathcal{V}$, we denote with $\mu(S)$ the *final active set*, i.e., the set of nodes that are active at the end of the diffusion process starting from S .

Upon the above defined quantities, we introduce a measure that will be essential to the definition of our targeted IM problem. This measure, which we denote as *delurking capital* cumulated via S , is proportional to the amount of the lurking scores over all target nodes that are activated by the seed set S .

Definition 1. Given a set $S \subseteq \mathcal{V}$, the **delurking capital** $DC(\mu(S))$ associated with the final active set $\mu(S) \subseteq \mathcal{V}$ is defined as:

$$DC(\mu(S)) = \sum_{\substack{v \in \mu(S) \setminus S \wedge \\ \ell(v) \geq LS}} \ell(v) \quad (1)$$

Note that in Eq. (1) we do not consider nodes that belong to the seed set S . The ratio behind this choice is that the selection of the seed set should not be biased by nodes with highest lurking scores; this will be made clear next in the definition of our objective function, which in fact relies on $DC(\mu(S))$.

Definition 2. Delurking-oriented Targeted Influence Maximization Problem. Given $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, b, \ell \rangle$, a diffusion model on \mathcal{G} , a budget k , and a lurking threshold LS , find a seed set $S \subseteq \mathcal{V}$ with $|S| \leq k$ of nodes (users) such that, by activating them, the final active set $\mu(S) \subseteq \mathcal{V}$ will have the maximum delurking capital:

$$S = \underset{S' \subseteq \mathcal{V} \text{ s.t. } |S'| \leq k}{\operatorname{argmax}} DC(\mu(S')) \quad (2)$$

The problem in Def. 2 preserves the complexity of the IM problem and, as a result, it is computationally intractable, i.e., it is still NP-hard. However, as for the classic IM problem, a greedy solution can be designed provided that the natural diminishing property holds for the considered problem. It is well-known that for many diffusion models, including LT, the function $\sigma(A)$ mapping any subset $A \subseteq \mathcal{V}$ of nodes to the size of the final active set $\mu(A)$ satisfies monotonicity and submodularity by exploiting the *equivalent live-edge graph model* [6].

Proposition 1. The delurking capital function DC defined in Eq. (1), mapping any active set $\mu(A) \subseteq \mathcal{V}$ to its overall delurking capital, is monotone and submodular, for any $LS \in [0, 1]$, under the LT model.

The above theoretical result enables the development of an approximate algorithm with guarantee for our problem, as discussed later in Sect. 2.3.

2.2 Choosing the information diffusion model

The LT model deals with situations in which exposure to multiple sources is needed for a user before taking a decision. More specifically, a node v can be influenced by each neighbor u according to a weight $b(u, v)$ such that $\sum_{u \in N^{in}(v)} b(u, v) \leq 1$, where $N^{in}(v)$ is the in-neighbor set of node v . At the very beginning of the diffusion process, each node v is assigned a threshold uniformly at random from $[0, 1]$. This represents the weighted fraction of v 's neighbors that must become active in order for v to become active itself. Intuitively, the higher is this threshold, the harder will be the task of enrolling v in a new trend, since the total influence weight must exceed its threshold.

We chose LT for modeling the process of influence propagation to pursue the goal of engaging lurkers. Unlike IC or IC-based models, the ability of LT to reflect the cumulative effect of exposure to multiple sources of influence can be profitably exploited to maximize the likelihood of changing the status of a user into a more active role in the OSN. Of course, this can actually lead to delurking provided that the information to be diffused towards the target lurkers is of any type that concerns a well-established delurking action [9]. In this regard, note that our approach is general as it can involve any type of delurking action in the form of piece of information to flow through the network (e.g., informative rewards, welcome messages, guidelines from experts, etc).

2.3 The DEVOTION algorithm

Our proposed greedy method, named DEVOTION (stands for **DE**lurking **O**riented **T**argeted **I**nfluence maximizati**ON**), exploits the search for shortest paths in the diffusion graph, following the lead of the study in [3], however in a backward fashion. Along with the information diffusion graph \mathcal{G} , the budget integer k , and the minimum lurking score LS , DEVOTION takes in input a real-valued threshold η . This parameter is used to control the size of the neighborhood within which paths are enumerated: in fact, the majority of influence can be captured by exploring the paths within a relatively small neighborhood; for higher η values, less paths are explored (i.e., paths are pruned earlier) leading to smaller runtime but with decreased accuracy in spread estimation.

The key idea of DEVOTION is to perform a backward visit of the graph starting from the nodes identified as target (i.e., the nodes u with $\ell(u) \geq LS$). In order to yield a seed set S of size at most k , DEVOTION works as follows. During each iteration, DEVOTION computes the set (say T) of nodes that reach the target ones and keeps track of the node with the highest marginal gain (i.e., delurking capital DC). The former allows an efficient reset of the spread of each non-seed node contained in T . The latter is found at the end of each iteration upon backward visit of all nodes in *TargetSet* that do not belong to the current seed set S . This visiting subroutine takes a path \mathcal{P} , its probability pp , and the lurking score of the end node in the path (i.e., a target node), and extends \mathcal{P} as much as possible (i.e., as long as pp is not lower than η). Initially, a path is formed by one target node, with probability 1; then, the path is extended by exploring the graph backward, adding to it one, unexplored in-neighbor v at time, in a depth-first fashion. The path probability is updated according to the LT-equivalent “live-edge” model [3,6], and so the delurking capital. The process is continued until no more nodes can be added to the path.

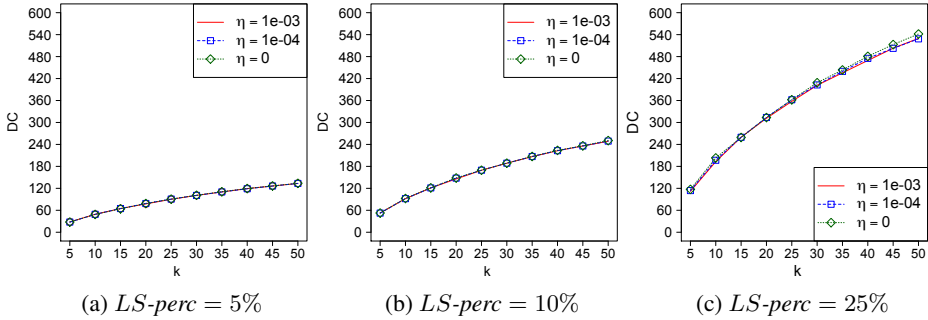


Fig. 1. Delurking capital in function of k and η on Instagram.

3 Experimental Evaluation

3.1 Evaluation methodology

We assessed the sensitivity of DEVOTION to the various parameters, i.e., the size of seed set (k), the lurking threshold (LS) for the selection of the target set, and the path pruning threshold (η). We also comparatively evaluated DEVOTION with SimPath [3] and TIM+ [7], two well-known state-of-the-art algorithms for IM problems, and with KB-TIM [13], a query-based targeted IM algorithm.

We used *FriendFeed* (493K, 19MM edges), *GooglePlus* (108K nodes, 14MM edges) and *Instagram* (18K nodes, 618K edges) network datasets, which are publicly available [11,12]. All experiments were carried out on an Intel Core i7-3960X CPU@3.30GHz, 64GB RAM machine.

3.2 Results

Sensitivity to parameters. We evaluated the performance of DEVOTION in terms of delurking capital obtained by varying all three parameters involved, i.e., k , LS -perc, η . We initially focused on η , by varying it from $1.0e-03$ to 0; note that $\eta = 1.0e-03$ is the default value used in other IM algorithms (e.g.,[3]), while $\eta = 0$ indicates no path pruning. We generally observed that no significant gain in DC is obtained for lower values of η . This fact achieves particular importance when it is coupled with the time performance results shown in Fig. 2: several orders of magnitude in the runtime are saved when $\eta > 0$. Within this view, $\eta = 1.0e-03$ might be preferred to $\eta = 1.0e-04$ when trade-off between DC performance and runtime has to be ensured.

Concerning the other two parameters, as expected, the delurking capital increases with both the size of the seed set (k) and the size of the target set; we controlled the latter through a parameter LS -perc which corresponds to a percentage value that determines the setting of LS such that the selected target set corresponds to the top- LS -perc of the distribution of node lurking scores. An important remark is that on all datasets a significant fraction of delurking capital, for a given choice of LS -perc, can be achieved just using low k . This appears evident, especially for lower LS -perc, in Fig. 1 by observing the increasing slope of the line fitting the DC curves as LS -perc increases.

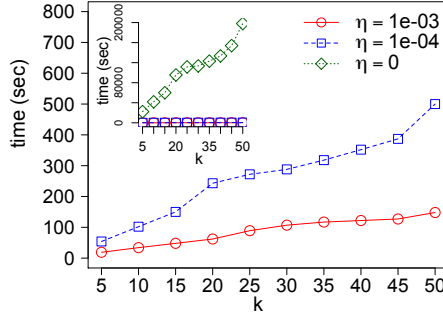


Fig. 2. Execution time of DEVOTION in function of k , with $LS\text{-perc} = 5\%$ and for varying η , on Instagram. The inset shows the overall results that include $\eta = 0$, while the main plot zooms in to focus on $\eta > 0$.

Concerning the other two datasets (results not shown), we observed that low values of k are enough to obtain a DC value close to or with the same order of magnitude of DC values achieved by using highest k .

Comparison with SimPath. SimPath shares with DEVOTION the kind of algorithmic approach to LT-based influence maximization, which includes the involvement of parameter η for controlling the enumeration of the paths through the diffusion graph. Results of the comparative evaluation for $\eta = 1.0e-04$, in terms of DC performances, are summarized in Table 1.

A first important remark that stands out is that the delurking capital obtained by SimPath is always lower than the one obtained by DEVOTION, for all datasets and configurations. The largest increment in DC is obtained for FriendFeed, with percentage of increment up to 10.8% (for $LS\text{-perc} = 5\%$). Significant increment is observed also w.r.t. the other datasets, with percentages of increment up to 4.27% (for $LS\text{-perc} = 5\%$) on GooglePlus, and up to 5.69% (for $LS\text{-perc} = 5\%$) on Instagram.

Impact of parameter η in the evaluation turns out to be marginal. Nevertheless, it should be noted that even though performance corresponding to $\eta = 1.0e-3$ (results not shown) are slightly lower in terms of DC , the increment w.r.t. SimPath is generally higher, indicating that DEVOTION is more robust than SimPath to path pruning.

As the size of target set increases, a general decreasing trend can be observed in the gap between DEVOTION and SimPath DC values. This is not surprising since a larger target set means a larger overlap with the entire node set, thus letting the seed nodes picked up by a non-targeted algorithm reach a larger part of the target set.

Regarding comparison in terms of running times, DEVOTION turns out to be two orders of magnitude faster than SimPath, for the same choice of η . Considering the largest dataset, FriendFeed, DEVOTION shows average running time of 47,278 seconds smaller than SimPath in its faster configuration ($\eta = 1.0e-03$, $LS\text{-perc} = 5\%$), and of 182,131 seconds in its slowest configuration ($\eta = 1.0e-04$, $LS\text{-perc} = 25\%$).

Comparison with TIM+. TIM+ follows an approach to influence maximization that was designed to improve efficiency in previously existing algorithms, including

Table 1. Summary of delurking capital performances of competitors against DEVOTION.

	FriendFeed			GooglePlus			Instagram		
	5%	10%	25%	5%	10%	25%	5%	10%	25%
incr. vs. SimPath	10.80%	9.32%	3.97%	4.27%	3.02%	1.06%	5.69%	4.10%	1.60%
incr. vs. TIM+	9.85%	8.82%	3.49%	4.22%	2.96%	1.01%	4.43%	3.11%	1.12%
incr. vs. KB-TIM	58.87%	34.48%	35.23%	40.66%	36.07%	33.86%	5.44%	6.79%	10.20%

SimPath. Therefore, it was not surprising to find out that TIM+ is indeed faster than DEVOTION. Taking FriendFeed as case in point, there is an average saving of 204 seconds for the fastest configurations ($\eta = 1.0e-03$ and $LS\text{-perc} = 5\%$ for DEVOTION, $\epsilon = 1.0$ for TIM+), and of 6100 seconds for the slowest configurations ($\eta = 1.0e-04$ and $LS\text{-perc} = 25\%$ for DEVOTION, $\epsilon = 0.1$ for TIM+).

Nevertheless, DEVOTION always outperforms TIM+ in terms of DC on all datasets and for all configurations. The performance gain achieved by DEVOTION is more significant on FriendFeed, with average percentage of increment of 9.85% (for $LS\text{-perc} = 5\%$), 8.82% ($LS\text{-perc} = 10\%$) and 3.49% ($LS\text{-perc} = 25\%$). Analogously to what we observed for the evaluation with SimPath and for the same reason, the gap in performance between DEVOTION and TIM+ decreases for larger target sets. Smaller but significant increment is also obtained on GooglePlus (average of 4.22% for $LS\text{-perc} = 5\%$) and Instagram (average of 4.43% for $LS\text{-perc} = 5\%$).

Comparison with KB-TIM. DEVOTION always performs significantly better than KB-TIM in terms of DC , on all datasets and with all configurations. Surprisingly, the gain in performance between DEVOTION and KB-TIM is larger than the one observed for other non-targeted competitors (i.e., SimPath, TIM+). Again, as previously seen for the comparative evaluation with IM algorithms, the highest increment corresponds to FriendFeed. As shown in Table 1, the percentage of average increment in DC ranges from 35.23% (for $LS\text{-perc} = 25\%$) to 58.87% (for $LS\text{-perc} = 5\%$). The increment is from 33.86% (for $LS\text{-perc} = 25\%$) to 40.66% (for $LS\text{-perc} = 5\%$) for GooglePlus, and from 5.44% (for $LS\text{-perc} = 5\%$) to 10.20% (for $LS\text{-perc} = 25\%$) for Instagram.

It should be noted that the comparison between DEVOTION and KB-TIM on Instagram is the only one following an opposite trend, w.r.t. other datasets and competitors, with the size of the target set: larger increments are achieved for larger target sets. This is probably due to the structural characteristics of Instagram and it is confirmed by a greater overlap between the seed sets of DEVOTION and KB-TIM than the seed overlap observed for the other datasets, as we shall discuss in the next section. As regards the execution times, KB-TIM performs comparably to TIM+, thus running always faster than DEVOTION and SimPath.

3.3 Discussion

Empirical evidence has shed light on the significance and effectiveness of DEVOTION. The method has shown to be robust w.r.t. the pruning of paths to be explored in the graph. A significant fraction of delurking capital can be achieved already with a small seed set, even for large network datasets. Compared with state-of-the-art IM and targeted IM algorithms, DEVOTION always obtains better quality seed sets, confirming its efficacy and uniqueness to solve a delurking-oriented targeted IM task.

4 Conclusions

We discussed the first computational approach to delurking, i.e., to engaging users that take on a lurking role in OSNs. We presented a novel targeted IM problem in which the objective function to be maximized is defined in terms of delurking capital of the target users. We proved that the proposed objective function is monotone and submodular, by using the LT-equivalent live-edge graph model, and developed an approximate algorithm, DEvOTION, to solve the problem under consideration.

The presented approach is independent on the particular strategy of delurking (being it based on rewards, welcome messages, etc.). We paved the way for the development of web-based systems that, by embedding our approach, can effectively support engagement of users. Also, the presented approach can easily be generalized to deal with other targeted IM scenarios, therefore we envisage further developments from various perspectives, including human-computer interaction and marketing.

References

1. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.A.: The role of social networks in information diffusion. In: Proc. World Wide Web Conf. (WWW), pp. 519–528 (2012)
2. Edelman, N.: Reviewing the definitions of “lurkers” and some implications for online research. *Cyberpsychology, Behavior, and Social Networking* **16**(9), 645–649 (2013)
3. Goyal, A., Lu, W., Lakshmanan, L.V.S.: SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model. In: Proc. IEEE Int. Conf. on Data Mining (ICDM), pp. 211–220 (2011)
4. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. *SIGMOD Record* **42**(2), 17–28 (2013)
5. Interdonato, R., Pulice, C., Tagarelli, A.: The DEvOTION algorithm for delurking in social networks. *Trends in Social Network Analysis, LNSN*, Springer, 28 pages (2017)
6. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: Proc. ACM Conf. on Knowledge Discovery and Data Mining (KDD), pp. 137–146 (2003)
7. Li, Y., Zhang, D., Tan, K.: Real-time targeted influence maximization for online advertisements. *PVLDB* **8**(10), 1070–1081 (2015)
8. Preece, J.J., Nonnecke, B., Andrews, D.: The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior* **20**(2), 201–223 (2004)
9. Sun, N., Rau, P.P.L., Ma, L.: Understanding lurkers in online communities: a literature review. *Computers in Human Behavior* **38**, 110–117 (2014)
10. Tagarelli, A., Interdonato, R.: “Who’s out there?”: Identifying and Ranking Lurkers in Social Networks. In: Proc. Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM), pp. 215–222 (2013)
11. Tagarelli, A., Interdonato, R.: Lurking in social networks: topology-based analysis and ranking methods. *Social Netw. Analys. Mining* **4**(230), 27 (2014)
12. Tagarelli, A., Interdonato, R.: Time-aware analysis and ranking of lurkers in social networks. *Social Netw. Analys. Mining* **5**(1), 23 (2015)
13. Tang, Y., Xiao, X., Shi, Y.: Influence maximization: near-optimal time complexity meets practical efficiency. In: Proc. ACM Conf. on Management of Data (SIGMOD), pp. 75–86 (2014)
14. Weng, L., Ratkiewicz, J., Perra, N., Gonçalves, B., Castillo, C., Bonchi, F., Schifanella, R., Menczer, F., Flammini, A.: The role of information diffusion in the evolution of social networks. In: Proc. ACM Conf. on Knowledge Discovery and Data Mining (KDD), pp. 356–364 (2013)