

# A Note on Maximizing the Spread of Influence in Social Networks

Eyal Even-Dar<sup>1</sup> and Asaf Shapira<sup>2</sup>

<sup>1</sup> Google Research

evendar@google.com

<sup>2</sup> Microsoft Research

asafico@microsoft.com

**Abstract.** We consider the *spread maximization* problem that was defined by Domingos and Richardson [6,15]. In this problem, we are given a social network represented as a graph and are required to find the set of the most “influential” individuals that by introducing them with a new technology, we maximize the expected number of individuals in the network, later in time, that adopt the new technology. This problem has applications in viral marketing, where a company may wish to spread the rumor of a new product via the most influential individuals in popular social networks such as Myspace and Blogosphere.

The spread maximization problem was recently studied in several models of social networks [10,11,13]. In this short paper we study this problem in the context of the well studied probabilistic *voter model*. We provide very simple and efficient algorithms for solving this problem. An interesting special case of our result is that the most natural heuristic solution, which picks the nodes in the network with the highest degree, is indeed the optimal solution.

## 1 Introduction

With the emerging Web 2.0, the importance of social networks as a marketing tool is growing rapidly and the use of social networks as a marketing tool spans diverse areas, and has even been recently used by the campaigns of presidential candidates in the United States. Social networks are networks (i.e. graphs) in which the nodes represent individuals and the edges represent relations between them. To illustrate the viral marketing channel (see [2,3,6]), consider a new company that wishes to promote its new specialized search engine. A promising way these days would be through popular social network such as Myspace, Blogosphere etc, rather than using classical advertising channels. By convincing several key persons in each network to adopt (or even to try) the new search engine, the company can obtain an effective marketing campaign and to enjoy the diffusion effect over the network. If we assume that “convincing” each key person to “spread” the rumor on the new product costs money, then a natural problem is the following: given a social network, how can we detect the players through which we can spread, or “diffuse”, the new technology in the most effective way.

Diffusion processes in social network have been studied for a long time in social sciences, see e.g. [5,8,3,16]. The algorithmic aspect of marketing in social networks was introduced by Domingos and Richardson [6,15] and can be formulated as follows. Given a social network structure and a diffusion dynamics (i.e. how the individuals influence each other), find a set  $S$  of nodes of cost at most  $K$  that by introducing them with a new technology/product, the spread of the technology/product will be maximized. We refer to the problem of finding such a maximizing set  $S$  as the *Spread maximization set problem*. The work of Domingos and Richardson [6,15] studied this problem in a probabilistic setting and mainly provided heuristics to compute a maximizing set. Following [6,15], Kempe et al. [10,11] and Mossel and Roch [13] considered a threshold network, in which users adopt a new technology only if a fixed fraction of their neighbors have already adopted this new technology. Their results show that finding the optimal subset of size  $K$  is NP-Hard to approximate within a factor smaller than  $1 - 1/e$  and also show that a greedy algorithm achieves this ratio.

**Our contribution:** In this paper we consider the Spread maximization set problem, in the case where the underlying social network behaves like the *voter model*. The voter model, which was introduced by Clifford and Sudbury [4] and Holley and Liggett [9], is probably one of the most basic and natural probabilistic models to represent the diffusion of opinions in a social network; it models the diffusion of opinions in a network as follows: in each step, each person changes his opinion by choosing one of his neighbors at random and adopting the neighbor's opinion. The model has been studied extensively in the field of interacting particle systems [12,1].

While the voter model is different from the threshold models that were studied in [10,11,13], it still has the same key property that a person is more likely to change his opinion to the one held by most of his neighbors. In fact, the threshold models of [10,11,13] are monotone in the sense that once a vertex becomes “activated” it stays activated forever. This makes these models suitable for studying phenomena such as infection processes. However, some process, such as which product a user is currently using, are not monotone in this sense. Therefore, the voter model, which allows to deactivate vertices, may be more suitable for studying non monotone processes.

Our main contributions are an exact solution to the spread maximization set problem in the voter model, when all nodes have the same cost (the cost of a node is the cost of introducing the person with a new technology/product), and providing an FPTAS<sup>1</sup> for the more general case in which different nodes may have different costs. In contrast to most of the previous results, which considered only the status of the network in the “limit”, that is, when the network converges to a steady state, our algorithms easily adopt to the case of different target

---

<sup>1</sup> An FPTAS, short for Fully Polynomial Time Approximation Scheme, is an algorithm that for any  $\epsilon$  approximates the optimal solution up to an error  $(1 + \epsilon)$  in time  $\text{poly}(n/\epsilon)$ .

times.<sup>2</sup> An interesting special case of our result is that the most natural heuristic solution, which picks the nodes in the network with the highest degree, is indeed the optimal solution, when all nodes have the same cost. We show that the optimal set for the long term is the set that maximizes the chances of reaching consensus with new technology/product.

## 2 The Voter Model

We start by providing a formal definition of the voter model (see [4,9] for more details).

**Definition 1.** Let  $G = G(V, E)$  be an undirected graph with self loops. For a node  $v \in V$ , we denote by  $N(v)$  the set of neighbors of  $v$  in  $G$ . Starting from an arbitrary initial 0/1 assignment to the vertices of  $G$ , at each time  $t \geq 1$ , each node picks uniformly at random one of its neighbors and adopts its opinion. More formally, starting from any assignment  $f_0 : V \rightarrow \{0, 1\}$ , we inductively define

$$f_{t+1}(v) = \begin{cases} 1, & \text{with probability } \frac{|\{u \in N(v) : f_t(u) = 1\}|}{|N(v)|} \\ 0, & \text{with probability } \frac{|\{u \in N(v) : f_t(u) = 0\}|}{|N(v)|} \end{cases}$$

Note that the voter model is a random process whose behavior depends on the initial assignment  $f_0$ . If we think of  $f_t(v) = 1$  as indicating whether  $v$  is using the product we wish to advertise, then a natural quantity we wish to study is the expected number of nodes satisfying  $f_t(v) = 1$  at any given time  $t$ . Of course, a simple way to maximize the number of such nodes is to start from an initial assignment  $f_0$  in which  $f_0(v) = 1$  for all  $v$ . However, in reality we may not be able to start from such an assignment as there is a cost  $c_v$  for setting  $f_0(v) = 1$  and we have a limited budget  $B$ . For example,  $c_v$  can be the cost of “convincing” a website to use a certain application we want other websites to use as well. This is the main motivation for the spread maximization set problem that is defined below in the context of the voter model. As we have previously mentioned, this (meta) problem was first defined by Domingos and Richardson [6,15] and was studied by [15,10,11,13] in other models of social networks.

**Definition 2 (The spread maximization set problem).** Let  $G$  be a graph representing a social network,  $\bar{c} \in \mathbb{R}^n$  a vector of costs indicating the cost  $c_v$  of setting  $f_0(v) = 1$ ,  $B$  a budget, and  $t$  a target time. The spread maximization set problem is the problem of finding an assignment  $f_0 : V \rightarrow \{0, 1\}$  that will maximize the expectation  $\mathbb{E} [\sum_{v \in V} f_t(v)]$  subject to the budget constraint  $\sum_{\{v : f_0(v) = 1\}} c_v \leq B$ .

<sup>2</sup> Kempe et al. [10] considered also finite horizon but under different objective function, i.e. for every individual how many timesteps she held the desired opinion until the target time. Furthermore, their approach required maintaining a graph whose size is proportional to the original graph size times the target time.

### 3 Solving the Spread Maximization Set Problem

Our algorithms for solving the spread maximization set problem all rely on the well known fact that the voter model can be analyzed using *graphical models* (see [7] for more details). Let us state a very simple yet crucial fact regarding the voter model that follows from this perspective. Recall that in the voter model, the probability that node  $v$  adopts the opinion of one of its neighbors  $u$  is precisely  $1/N(v)$ . Stated equivalently, this is the probability that a random walk of length 1 that starts at  $v$  ends up in  $u$ . Generalizing this observation to more than one step, one can easily prove the following by induction on  $t$ .

**Proposition 1.** *Let  $p_{u,v}^t$  denote the probability that a random walk of length  $t$  starting at node  $u$  stops at node  $v$ . Then the probability that after  $t$  iterations of the voter model, node  $u$  will adopt the opinion that node  $v$  had at time  $t = 0$  is precisely  $p_{u,v}^t$ .*

We thus get the following corollary.

**Corollary 1.** *Let  $S = \{u : f_0(u) = 1\}$ . The probability that  $f_t(v) = 1$  is the probability that a random walk of length  $t$  starting at  $v$  ends in  $S$ .*

Equipped with the above facts we can now turn to describe the simple algorithms for the spread maximization set problem.

**The case of short term:**

We start by showing how to solve the problem for the case of the short term, that is when  $t$  is (any) polynomial in  $n$ . We note that studying the spread maximization problem for short time term is crucial to the early stages of introducing a new technology into the market. As usual, let  $M$  be the normalized transition matrix of  $G$ , i.e.  $M(v, u) = 1/|N(v)|$ . For a subset  $S \subseteq \{1, \dots, n\}$  we will denote by  $1_S$  the 0/1 vector, whose  $i^{th}$  entry is 1 iff  $i \in S$ . The following lemma gives a characterization of the spread maximizing set.

**Lemma 1.** *For any graph  $G$  with transition matrix  $M$ , the spread maximizing set  $S$  is the set which maximizes  $1_S M^t$  subject to  $\sum_{v \in S} c_v \leq B$ .*

*Proof.* Recall the well known fact that  $p_{u,v}^t$ , which is the probability that a random walk of length  $t$  starting at  $u$  ends in  $v$ , is given by the  $(u, v)$  entry of the matrix  $M^t$ . The spread maximizing set problem asks for maximizing  $\mathbb{E} [\sum_{v \in V} f_t(v)]$  subject to  $\sum_{v \in S} c_v \leq B$ . By linearity of expectation, we have that  $\mathbb{E} [\sum_{v \in V} f_t(v)] = \sum_{v \in V} Prob[f_t(v) = 1]$ . By Corollary 1 we have that if we set  $f_0(v) = 1$  for any  $v \in S$  then  $Prob[f_t(v) = 1] = 1_S M^t 1_{\{v\}}^T$ . Therefore,  $\mathbb{E} [\sum_{v \in V} f_t(v)] = \sum_{v \in V} 1_S M^t 1_{\{v\}}^T = 1_S M^t$ , and we conclude that the optimal set  $S$  is indeed the one maximizing  $1_S M^t$  subject to  $\sum_{v \in S} c_v \leq B$ . ■

Using this formulation we can obtain the following theorems that shed light on how well can be the maximizing spread set problem solved.

**Theorem 1.** *If the vector cost  $\bar{c}$  is uniform, that is, if for all  $v$  we have  $c_v = c$ , then the spread maximization set problem can be solved exactly in polynomial time for any  $t = \text{poly}(n)$ .*

*Proof.* First note the entries of  $M^t$  can be computed efficiently for any  $t = \text{poly}(n)$ . For any  $t$  to compute  $M^t$  we need to perform  $O(\log t)$  matrix multiplication which can be done efficiently. For every node  $v$  denote  $g_v = 1_{\{v\}}M^t$ . By Lemma 1 we have that the problem is equivalent to the problem of maximizing  $1_S M^t$  subject to  $\sum_{v \in S} c_v \leq B$ . As  $1_S M^t = \sum_{v \in S} g_v$  and the cost of every node is identical, we get that for every budget  $B$ , the optimal set is the first  $\lfloor B/c \rfloor$  nodes when sorted according to  $g_v$ . ■

**Theorem 2.** *There exists an FPTAS to the spread maximization set problem for any  $t = \text{poly}(n)$ .*

Observe that in general we cannot expect to be able to solve the spread maximization set problem exactly because when  $t = 0$  this problem is equivalent to the Knapsack problem, which is NP-hard.

**The case of long term:**

Let us consider now the case of large  $t$ , where by large we mean  $t \geq n^5$ . Recall the well known fact that for any graph  $G$  with self loops, a random walk starting from any node  $v$ , converges to the steady state distribution after  $O(n^3)$  steps (see [14]). Furthermore, if we set  $d_v = |N(v)|$  then the (unique) steady state distribution is that the probability of being at node  $u$  is  $d_u/2|E|$ . In other words, if  $t \gg n^3$  then  $M_{u,v}^t = (1 + o(1))d_u/2|E|$ .<sup>3</sup> Once again, using Lemma 1 we can obtain the following theorem.

**Theorem 3.** *There exists a linear time FPTAS to the spread maximization set problem when  $t \geq n^5$ .*

An interesting special case of Theorem 3 is when all nodes have the same cost  $c$ . Observe that in this case we get that the optimal solution is simply to pick the  $\lfloor B/c \rfloor$  vertices of  $G$  of highest degree. This gives a formal justification for the “heuristic” approach of picking the nodes in the social network with the largest number of acquaintances, e.g. [17,6,15].

**Maximizing the probability of consensus:**

It is a well known fact that after  $O(n^3 \log n)$  time the voter model reaches a consensus with high probability, that is, when  $t \geq n^3 \log n$  either  $f_t(v) = 1$  for all  $v$  or  $f_t(v) = 0$  for all  $v$ .

**Theorem 4.** *With probability  $1 - o(1)$ , the voter model converges to consensus after  $O(n^3 \log n)$  steps.*

By Theorems 3 and 4 we derive the following corollary,

**Corollary 2.** *For any  $t \geq n^3 \log n$  and  $\epsilon > 0$ , there is a linear time algorithm for maximizing, up to an additive error of  $\epsilon$ , the probability that the voter model reaches an all-ones consensus after  $t \geq n^3 \log n$  steps.*

<sup>3</sup> More precisely, the smaller we want the  $o(1)$  term to be the larger we need  $t$  to be.

**Acknowledgments.** The authors would like to thank Michael Kearns and Yuval Peres for valuable discussions concerning the voter model.

## References

1. Aldous, D., Fill, J.: Reversible Markov Chains and Random Walks on Graphs (Draft). Draft (2007)
2. Bass, F.M.: A new product growth model for consumer durables. *Management Science* 15, 215–227 (1969)
3. Brown, J.J., Reinegen, P.H.: Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*
4. Clifford, P., Sudbury, A.: A model for spatial conflict. *Biometrika* 60(3), 581–588 (1973)
5. Coleman, J.S., Katz, E., Menzel, H.: *Medical Innovations: A Diffusion Study*. Bobbs Merrill (1966)
6. Domingos, P., Richardson, M.: Mining the network value of customers. In: *KDD*, pp. 57–66 (2001)
7. Durrett, R.: *Lecture Notes on Particle Systems and Percolation*. Wadsworth (1988)
8. Granovetter, M.: Threshold models of collective behavior. *American Journal of Sociology* 83(6), 1420–1443 (1978)
9. Holley, R.A., Liggett, T.M.: Ergodic theorems for weakly interacting infinite systems and the voter model. *Annals of Probability* 3, 643–663 (1975)
10. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: *KDD*, pp. 137–146 (2003)
11. Kempe, D., Kleinberg, J.M., Tardos, E.: 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)
12. Liggett, T.M.: *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Springer, Heidelberg (1999)
13. Mossel, E., Roch, S.: On the submodularity of influence in social networks. In: *STOC*, pp. 128–134 (2007)
14. Motwani, R., Raghavan, P.: *Randomized Algorithms*. Cambridge University Press, Cambridge (1996)
15. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: *KDD*, pp. 61–70 (2002)
16. Valente, T.: *Network Models of the Diffusion of Innovations*. Hampton Press (1995)
17. Wasserman, S., Faust, K.: *Social Network Analysis*. Cambridge University Press, Cambridge (1994)