A Note on Maximizing the Spread of Influence in Social Networks

Eyal Even-Dar[1] and Asaf Shapira[2]

[1] Google Research, Email: evendar@google.com
[2] Microsoft Research, Email: asafico@microsoft.com

Abstract. We consider the *spread maximization* problem that was defined by Domingos and Richardson [7, 22]. 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 Blogsphere.
The spread maximization problem was recently studied in several models of social networks [14, 15, 20]. 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

1.1 Background

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[1]. 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, 7, 11, 19]), 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, Blogsphere 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, 12, 3, 23, 24]. The algorithmic aspect of marketing in social networks was

[1] Hillary Clinton - http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendID=64552165, Barack Obama - http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=184040237, John Edwards-http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=9736082, Rudy Giuliani - http://www.myspace.com/rudygiulianiisgod

introduced by Domingos and Richardson [7, 22] 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 [7, 22] studied this problem in a probabilistic setting and mainly provided heuristics to compute a maximizing set. Following [7, 22], Kempe et al. [14, 15] and Mossel and Roch [20] 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.

1.2 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 [13], 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 [17, 18, 10, 1] and many variations of the network structure have been analyzed, e.g. d-dimensional integer lattice [4, 13], finite torus [6], finite graphs [8], regular graphs [1] and small world graphs [10].

While the voter model is different from the threshold models that were studied in [14, 15, 20], 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 [14, 15, 20] 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. Another important property of the voter model is that a consensus is reached with probability one (see Theorem 4). It is interesting to observe that many technologies (almost) reach consensus, for instance Windows as an operating system, Google as a search engine, YouTube for sharing videos and many more.

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 [2] for the more general case in which different nodes may have different costs. In contrast to most of the

[2] An FPTAS, short for Fully Polynomial Time Approximation Scheme, is an algorithm that for any ϵ approximates the optimal solution up to an error $(1 + \epsilon)$ in time poly(n/ϵ).

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 times. [3] 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.

We note that while our results assume a synchronous model, i.e. at each step all the users are updating their opinions, and unweighted graph all the results apply to asynchronous models and weighted graphs with very simple modification that are omitted from this extended abstract.

2 The Voter Model

We start by providing a formal definition of the voter model (see [4, 13] 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 $f_0 : V \to \{0, 1\}$, 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 previous mentioned, this (meta) problem was first defined by Domingos and Richardson [7, 22] and was studied by [22, 14, 15, 20] 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 $f_0 : V \to \{0, 1\}$ that will maximize the expectation $\mathbb{E}\left[\sum_{v \in V} f_t(v)\right]$ subject to the budget constraint $\sum_{\{v : f_0(v) = 1\}} c_v \leq B$.*

[3] Kempe et al. [14] 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 [9] 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 $p^t_{u,v}$ 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 $p^t_{u,v}$.*

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.

3.1 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 $S \subseteq \{1, \ldots, n\}$ 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 $p^t_{u,v}$, 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 $\mathbb{E}\left[\sum_{v \in V} f_t(v)\right]$ subject to $\sum_{v \in S} c_v \leq B$. By linearity of expectation, we have that

$$\mathbb{E}\left[\sum_{v \in V} f_t(v)\right] = \sum_{v \in V} Prob[f_t(v) = 1].$$

By Corollary 1 we have that if we set $f_0(v) = 1$ for any $v \in S$ then

$$Prob[f_t(v) = 1] = 1_S M^t 1^T_{\{v\}} .$$

Therefore,

$$\mathbb{E}\left[\sum_{v\in V} f_t(v)\right] = \sum_{v\in V} 1_S M^t 1_{\{v\}}^T = 1_S M^t \ ,$$

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. We note that these positive results are in contrast to the inapproximabilty results in the model introduced by [14] for threshold networks.

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 $t = poly(n)$.*

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 preform $O(\log t)$ 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 $t = poly(n)$.*

Proof. Once again, for every node v denote $g_v = 1_{\{v\}} M^t$. Our goal is to maximize $1_S M^t = \sum_{v\in S} g_v$ subject to $\sum_{v\in S} p_v \leq B$. Observe that this is just an instance of the Knapsack problem and thus we can use the well known linear time FPTAS algorithm of Knapsack [16] to obtain an FPTAS to the spread maximization set problem. ■

Observe that in general we can 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.

3.2 The case of long term

In the previous subsection we have considered the case where t is polynomial in n. 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 [21]). 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 $M^t_{u,v} = (1 + o(1))d_u/2|E|$. [4] Once again, using Lemma 1 we can obtain the following corollaries.

[4] More precisely, the smaller we want the $o(1)$ term to be the larger we need t to be.

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

Proof. If $t \geq n^5$ then by the above observation we know the approximate entries of M^t without actually computing M^t. The error in each entry is within a factor of $1 + o(1/n^2)$ of the exact value, so we can use the linear time FPTAS of Knapsack [16] as in Theorem 2. ■

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. [25, 7, 22].

3.3 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 $f_t(v) = 1$ for all v. Let us sketch the simple proof of this fact for completeness.

Theorem 4. *With probability $1-o(1)$, the voter model converges to consensus after $O(n^3 \log n)$ steps* [5].

Proof. (sketch) Recall that by Lemma 1 the opinion of node v in time t is distributed according to a random walk starting at v of length t. Now, for every pair of vertices u and v, it is well known (see [1]) that with probability $1 - o(1/n)$ two random walks starting at u and v will meet, with probability $1 - o(1/n)$ after $n^3 \log n$ steps. This means, that with probability $1 - o(1/n)$ vertices u and v will have the same value. By the union bound, we conclude that after $O(n^3 \log n)$ steps all the vertices will have the same value with probability $1 - o(1)$. ■

By Theorem 4 we know that when $t \geq n^3 \log n$ then either all vertices hold the value 1 or none of them. We thus conclude that for $t \geq n^3 \log n$ the probability of reaching an "all-ones" consensus is $\frac{1}{n}\mathbb{E}\left[\sum_{v \in V} f_t(v)\right] - o(1)$. Since by Theorem 3 we can efficiently approximate the maximal value of $\mathbb{E}\left[\sum_{v \in V} f_t(v)\right]$ we get the following simple 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 ϵ, the probability that the voter model reaches an all-ones consensus after $t \geq n^3 \log n$ steps.*

Let us conclude by noting that the assumption of this subsection that $t \geq n^5$ was only needed in order to guarantee (with slackness) that a random walk on G converges to the stationary distribution after t steps. Of course, of the graph G has the property that random walks on it mix much faster (eg, when G is an expander graph) one can apply the algorithm described in this subsection already for much smaller values of t.

[5] See [1], Chapter 14, for more refined versions of this theorem.

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

References

1. D. Aldous and J. Fill. *Reversible Markov Chains and Random Walks on Graphs*. 2007. Draft.
2. F. M. Bass. A new product growth model for consumer durables. *Management Science*, 15:215–227, 1969.
3. J. J. Brown and P. H. Reinegen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*.
4. P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581–588, 1973.
5. J. S. Coleman, E. Katz, and H. Menzel. *Medical Innovations: A Diffusion Study*. Bobbs Merrill, 1966.
6. J. T. Cox. Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. *The Annals of Probability*, 17(4):1333–1366, 1989.
7. P. Domingos and M. Richardson. Miningthe network value of customers. In *Seventh International Conference on Knowledge scovery and Data Mining (KDD)*, pages 57–66, 2001.
8. P. Donnelly and D.J.A Welsh. Finite particle systems and infection models. *Math. Proc. Cambridge Philos. Soc*, 94:167–182, 1983.
9. R. Durrrett. *Lecture Notes on Particle Systems and Percolation*. Wadsworth, 1988.
10. R. Durrrett. *Random Graph Dynamics*. Cambridge University Press, 2007.
11. J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying processof word-of-mouth. *Marketing Letters*, 12:3:211–223, 2001.
12. M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.
13. R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite. systems and the voter model. *Annals of Probability*, 3:643–663, 1975.
14. D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *The Nineth International Conference on Knowledge scovery and Data Mining (KDD)*, pages 137–146, 2003.
15. D. Kempe, J. M. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *32nd International Colloquium on Automata, Languages and Programming(ICALP)*, pages 1127–1138, 2005.
16. E. L. Lawler. Fast approximation algorithm for knapsack problems. *Mathematics of Operations Research*, 4:339–356, 1979.
17. T. M. Liggett. *Interacting Particle Systems*. 1985.
18. T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Springer, 1999.
19. V. Mahajan, E. Muller, and F. M. Bass. New product diffusion models in marketing: A reviewand directions for research. *Journal of Marketing*, 54:1:1–26, 1990.
20. E. Mossel and S. Roch. On the submodularity of influence in social networks. In *39th Annual ACM Symposium on Theory of Computing(STOC)*, pages 128–134, 2007.
21. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1996.
22. M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Eighth International Conference on Knowledge scovery and Data Mining (KDD)*, pages 61–70, 2002.
23. E. M. Rogers. *Diffusion of innovations*. Free Press, 1995.
24. T. Valente. *Network Models of the Diffusion of Innovations*. Hampton Press, 1995.
25. S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.