Determining the Top-k Nodes in Social Networks using the Shapley Value

(Short Paper)

N. Rama Suri*
Electronic Commerce Laboratory,
Dept. of Computer Science and Automation,
Indian Institute of Science, Bangalore, India
nrsuri@csa.iisc.ernet.in

Y. Narahari†
Electronic Commerce Laboratory,
Dept. of Computer Science and Automation,
Indian Institute of Science, Bangalore, India
hari@csa.iisc.ernet.in

ABSTRACT

In this paper, we consider the problem of selecting, for any given positive integer k, the *top-k* nodes in a social network, based on a certain measure appropriate for the social network. This problem is relevant in many settings such as analysis of co-authorship networks, diffusion of information, viral marketing, etc. However, in most situations, this problem turns out to be NP-hard. The existing approaches for solving this problem are based on approximation algorithms and assume that the objective function is sub-modular. In this paper, we propose a novel and intuitive algorithm based on the *Shapley value*, for efficiently computing an approximate solution to this problem. Our proposed algorithm does not use the sub-modularity of the underlying objective function and hence it is a general approach. We demonstrate the efficacy of the algorithm using a co-authorship data set from e-print arXiv (www.arxiv.org), having 8361 authors.

Categories and Subject Descriptors

H.0 [**Information Systems**]: General; H.5.3 [**Group and Organization Interfaces**]: Evaluation/methodology

General Terms

Algorithms, Economics, Performance

Keywords

Social Networks, co-authorship networks, Shapley value, approximation algorithms

1. INTRODUCTION

A social network is a social structure made of individuals or organizations that are tied by one or more specific types of interdependency, such as friendship, co-authorship, collaboration, web links, etc. Typically, each individual is represented by a node in the network, and there is an edge between two nodes if there exists a social interaction between them [8]. Finding patterns of social interaction within a population has wide-ranging applications, which includes knowing the social and organizational structure. Given a social network, there has been quite intensive interest to find the *influential* nodes based on a well defined measure. We now present a few motivating examples in this regard.

Our first example is *diffusion of information* in social networks. In general, social networks play a key role for the spread of information or behavior within a population of individuals. The idea behind diffusion of diffusion is the extent to which people are likely to be influenced by the decisions of their neighbors. For a given value of k, a natural question that emerges is which subset of nodes with cardinality k should we target to maximize the size of the information cascade [2, 3] in a social network.

Our second example concerns analysis of collaboration patterns among research communities. There exists a natural social network here, where a node corresponds to a researcher and an edge exists between two researchers if they have collaborated (for example, co-authored in a paper). In such a network, it may be desirable to find the most prolific researchers since they are most likely the trend setters for new innovations.

Our third example is concerned with a social network of books sold by an online book seller such as Amazon.com (www.amazon.com). Edges between books represent the co-purchasing of books by the same buyers. It may be useful to find the most frequently co-purchased books because it would help the online book seller to enhance the quality of its recommendations to the users and also helps him to maximize his own profits.

In all the above mentioned contexts, the common goal is to find the influential nodes in the social network with respect to a measure that can capture the behavior in which we are interested. This task may turn out to be computationally hard in some contexts such as diffusion of information. Domingos [1] considered the problem of finding a set of nodes with cardinality k that can maximize the information cascade in viral marketing setting and proposed predictive models to show that selecting the right set of users for a marketing campaign can make a big difference. Later Kempe, Kleinberg, and Tardos [2] approached this problem from the

*N. Rama Suri is a Doctoral student in the Dept of Computer Science Automation, Indian Institute of Science, Bangalore. He is supported with MSR India PhD Fellowship.

†Y. Narahari is a Professor in the Dept of Computer Science and Automation, Indian Institute of Science, Bangalore.

Cite as: Determining the Top-k Nodes in Social Networks using the Shapley Value (Short Paper), N. Rama Suri and Y. Narahari, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16.,2008,Estoril,Portugal,pp. 1509-1512.

perspective of two widely studied operational models for diffusion of information, namely *thresholds model* and *cascade model*. They showed that the underlying objective function of the problem is NP-hard and further showed that it is sub-modular. They proposed provable approximation guarantees for this problem. They also presented a framework to generalize the thresholds model and the cascade model for reasoning about the performance guarantees of algorithms for these types of influence problems in social networks. In another paper, Kemp, Kleinberg, and Tardos [3] consider the problem of information diffusion in the presence of *word-of-mouth* referral and give a general model called the *decreasing cascade model*. The authors show that, in the presence of decreasing cascade model, a natural greedy algorithm achieves an approximation factor of $(1 - \frac{1}{e} - \epsilon)$ where $\epsilon > 0$. In both these papers, the authors point out that the greedy approximation algorithms suffer from a limitation that the underlying objective function for information diffusion can not be computed exactly and efficiently. Due to this reason, they efficiently simulate the diffusion process to determine approximate value of the underlying objective function. They further posed a question how to find influential nodes more efficiently in a systematic manner?

In this paper, we propose a novel and intuitive way to find the influential nodes with respect to a defined measure in social networks using the notion of the Shapley value, a well known concept in cooperative game theory [7, 4]. Our approach leads to an efficient algorithm for computing the k most influential nodes approximately.

For ease of explanation, we develop this approach in the context of coauthorship networks. In such a network, each node represents a researcher. There exists an edge between two nodes if the corresponding researchers have coauthored in a paper. Given a value for k, we need to find a set of k researchers who have coauthored with maximum number of other researchers. From now on we refer to this problem as *top-k nodes* problem in coauthorship networks. Let $N = \{1, 2, \ldots, n\}$ be the set of nodes in the network. For any $S \subseteq N$, we define a function $g(S)$ that represents the number of nodes that are adjacent to nodes in the set S. We call the function $g(.)$ as *top-k nodes function*. Given a value for k, we show that the problem of finding a set S of cardinality k such that $g(S)$ attains maximum value is NP-hard. This motivates us to think about efficiently finding a set of cardinality k that can approximate the optimal solution. Our contribution is to propose an intuitive algorithm based on the *Shapley value* for this purpose. It also turns out that our approach can be applied even when the underlying objective function is not sub-modular, as assumed in [2, 3]. We evaluate the the performance of the proposed algorithm using a co-authorship data set with 8361 authors from e-print arXiv (www.arxiv.org) and show that it outperforms a well known benchmark heuristic algorithm.

1.1 Organization of the Paper

The rest of the paper is organized as follows. In Section 2, we give a brief overview of Shapley value. In Section 3, we prove a few properties of the objective function involved in the top-k nodes problem and present our algorithm to find an approximate solution to the top-k nodes problem. In Section 4, we evaluate the performance of the proposed approach using a coauthorship data set. We finally conclude the paper in Section 5.

2. PRELIMINARIES

Here we present a brief discussion on cooperative game theory. A cooperative game with transferable utility is defined as the pair (N, v) where $N = \{1, 2, \ldots, n\}$ is the set of players and $v : 2^N \rightarrow \mathbb{R}$ is a characteristic function, that assigns a value to each subset of N, with $v(\phi) = 0$. The value $v(S)$ for a subset S of N (also called the *worth* of coalition S) represents the total utility that can be attained by the members in S only without any help from the members in $N \setminus S$.

The Shapley value [7, 4] is a solution concept for cooperative games which predicts a unique expected utility allocation for each player in the game. The Shapley value tries to capture how coalitional competitive forces influence the possible outcomes of a game. It describes a reasonable or fair way of dividing the gains from cooperation given the strategic realities captured by the characteristic function. It captures the *marginal contribution* that each player makes to the dynamics of the game. Given a cooperative game with transferable utilities, (N, v), the Shapley value, is the vector $\Phi(v) = (\Phi_1(v), \Phi_2(v), \ldots, \Phi_n(v))$, where

$$\Phi_i(v) = \frac{1}{n!} \sum_{\pi \in \Omega} [v(S_i(\pi) \cup i) - v(S_i(\pi))] \quad (1)$$

where Ω is the set of permutations over N, and $S_i(\pi)$ is the set of players appearing before the ith player in permutation π. The Shapley value $\Phi_i(v)$ of player i is the sum of marginal contributions of the player over all possible permutations averaged over the number of permutations.

3. THE PROPOSED APPROACH

Here we first present a few properties of the top-k nodes problem and then present our algorithm to this problem.

3.1 Properties of the Top-k Nodes Function

We now prove a few properties of the top-k node function $g(.)$.

Proposition 1: Given any co-authorship network, the top-k nodes function $g(.)$ is submodular.

Proof: Proof is an easy consequence of the definition $g(.)$.

Proposition 2: The top-k nodes function $g(.)$ is monotonically non-decreasing.

Proof: Directly follows from Proposition 1.

Proposition 3: Given a co-authorship network, the problem of finding the top-k nodes is NP-hard.

Proof: Let us consider the following instance of the k-coverage problem which is known to be NP-hard. Let U be the universal set with n elements. Each element $u \in U$ has an associated weight $w(u)$. Let $\mathcal{C} = \{A_1, A_2, \ldots, A_m\}$ be the set of subsets of U. We are given an integer k. We wish to know whether there exist k subsets of U, where each subset selected is a member of $\mathcal{C}$ such that the weight of the elements in $U_{i=1}^{k} A_i$ is maximized.

From the above instance of the k-coverage problem, we can construct an instance of the top-k nodes problem. Note that if any of the weights of the elements in U are negative or fractional, we can add appropriately a positive quantity to each of these elements and ensure that all the weights in U are positive integers. Let us call the new weight of an element u in U be $w^{'}(u)$. We define a new graph $G^{'}$

with $n + \sum_{i=1}^{n} w^{'}(i)$ nodes, i.e., there exists a node i corresponding to each subset A_i in $\mathcal{C}$ and there exist $w^{'}(u)$ nodes corresponding to each element u in U. If an element u belongs to set A_i, then there exist directed edges from i to the corresponding $w^{'}(u)$ nodes of u in $G^{'}$. If we pick the k nodes corresponding to the k sets in the solution of the k-coverage problem, it results in a solution to the top-k nodes problem. On the other hand, if we can pick k nodes that solve the top-k nodes problem, then we have a solution to the k-coverage problem.

It is clear that to address the top-k nodes problem, we want to find a set S of cardinality k such that $g(S)$ is maximized. Proposition 3 shows that it is an NP-hard problem. This motivates us to think about how to approximate the optimum solution for the top-k nodes problem. Nemhauser, Wolsey and Fisher [5] have presented a greedy hill climbing algorithm that achieves an approximation factor of $(1 - \frac{1}{e})$. Proposition 4 states this result more formally.

Proposition 4: For any non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let S^* be a set that maximizes the value of f over all k-element sets. Then $f(S) \geq (1 - \frac{1}{e}).f(S^*)$. That is, S achieves a $(1 - \frac{1}{e})$-approximation.

This greedy algorithm assumes that it is possible to evaluate the underlying function $f(.)$ exactly, which *may not* be the case in several contexts in social networks [2, 3]. To overcome the above serious difficulty, we propose an approach based on the Shapley value. The motivation for using the Shapley value arises from the fact that the Shapley value for each node (or player) gives the marginal contribution the node makes to the coalitional dynamics in the game. The higher the marginal contribution of a node, the higher the Shapley value of that node and the more *important* that node is among the players. It therefore makes sense to use the the Shapley value to pick an element with largest marginal gain in each iteration of the the greedy algorithm [5] as stated in Proposition 4. For this, we need to model the top-k nodes problem as an appropriate cooperative game.

The use of Shapley value in this current problem setting has another significant advantage, namely that the function $g(.)$ need not be submodular.

3.2 Shapley Value Based Algorithms

First, we model the top-k nodes problem as a cooperative game, (N, v). We define N to be the set of nodes in the co-authorship network. For each subset $S \subseteq N$, we define $v(S)$, in a natural way, as the number of nodes that are adjacent to the nodes in S.

We now outline a naive algorithm, based on Shapley value, for the top-k nodes problem.

Algorithm 1: Naive Algorithm

1. Consider the set, Ω, of $n!$ permutations of the nodes in the set N. Note that the size of each permutation is n.
2. For each permutation $(x_1, x_2, \ldots, x_n)$ in Ω, compute the marginal contribution of each node i using the following expression:
$$v(\{x_1, x_2, \ldots, x_i\}) - v(\{x_1, x_2, \ldots, x_{i-1}\})$$
3. Compute the Shapley value of each node in N using expression (1).
4. Pick k nodes with the highest Shapley values. The solution to the problem is the set consisting of these k nodes.

In *Algorithm 1*, it is clear that we have to compute Shapley value of each node while considering all possible $n!$ permutations of the nodes. It is easy to see that the running time of *Algorithm 1* is $O(\frac{n}{e})^n$. Thus this algorithm finds an approximate solution to the top-k nodes problem, but not in an efficient way. We propose to circumvent this difficulty by finding approximate Shapley values of the nodes in the co-authorship network in polynomial time. We do this by using a randomly sampled subset, call it Ψ, of permutations where the cardinality of Ψ is polynomial in n. Let t be the cardinality of Ψ, i.e., $t = |\Psi|$. We can compute the approximate Shapley values of nodes using the following algorithm.

Algorithm 2: Efficient Approximate Algorithm

1. Consider the set Ψ with t permutations of the nodes in N, where $t << n!$. Note that the size of each permutation is n.
2. For each permutation $(x_1, x_2, \ldots, x_n)$ in Ψ, compute the marginal contribution of each node i using the following expression:
$$v(\{x_1, x_2, \ldots, x_i\}) - v(\{x_1, x_2, \ldots, x_{i-1}\})$$
3. Compute the Shapley value of each node in Ψ using expression $\{\frac{1}{t!} \sum_{\pi \in \Psi} [v(S_i(\pi) \cup i) - v(S_i(\pi))\}$.
4. Pick k nodes with the highest Shapley values and the solution is given by these k nodes.

In *Algorithm 2*, we have to compute the marginal contribution of each node corresponding to each permutation in Ψ. This takes $O(tn)$ time. Picking k nodes with the k highest Shapley values takes $O(k\ log(n))$ time. The overall running time of *Algorithm 2* therefore is $O(tn + k\ log(n))$, and also we claim that t is polynomial in n.

Recall that we need to find a node, say v, with a high marginal gain in each iteration of the greedy algorithm [5] as stated in Proposition 4. We conjecture that Algorithm 2 picks such a node v in each iteration that is $(1 - \epsilon)$ approximate best node, where $\epsilon > 0$. Given that this conjecture is true, then by invoking *Theorem 1* in [3], we can claim that *Algorithm 2* achieves an approximation factor of $(1 - \frac{1}{e} - \epsilon^{'})$ where $\epsilon^{'}$ depends on ϵ polynomially.

4. EXPERIMENTS

In this section, we show the efficacy of *Algorithm 2* by conducting experiments on a real world co-authorship data set and comparing its performance against that of a well-known benchmark heuristic, the *maximum degree heuristic* [8]. as a baseline for all our comparisons. In applying the

maximum degree heuristic to address top-k nodes problem, we simply pick k nodes in the co-authorship network having the k highest degrees.

We construct a co-authorship network with 8361 researchers using the co-authorships in *high-energy Physics theory* publications. These co-authorships are between scientists posting preprints on the high-energy Physics theory e-print archive (www.arxiv.org) between Jan 1, 1995 and December 31, 1999. More information on this data set is available in [6]. In this co-authorship network, there is a node corresponding to each scientist and there exists a link between two scientists if they have co-authored at least one paper.

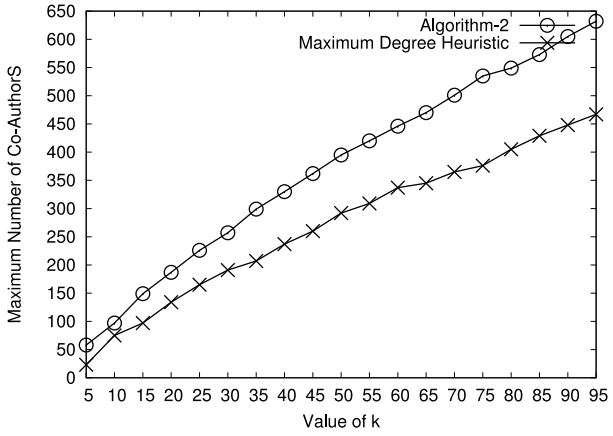

Figure 1: Shapley value based approach versus maximum degree heuristic based approach

The values that we consider for k are 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90. For each possible value of k, we compute the approximate solutions to the top-k nodes problem using *Algorithm 2* and using the maximum degree heuristic. These values are averaged over 1000 runs. Results shown in Figure 1 clearly indicate that the performance of Shapley value based algorithm superior than that of the maximum degree heuristic.

We now give a brief note on the size of the sampled set Ψ in *Algorithm 2*. Recall that there are $n = 8361$ researchers in the co-authorship network. So we have to sample a polynomial number of permutations in n into the set Ψ out of all possible 8361! (very huge number) permutations. We illustrate this sampling process for $k = 15$. We work with different sampled sets of sizes 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000. The graph in Figure 2 shows the maximum number of co-authors with each size of sampled set. This graph shows that we can get convincingly accurate results even with moderate sizes of sampled sets.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we considered the top-k nodes problem where we need to find k most influential nodes in the social network. We proposed an algorithm based on the Shapley value for efficiently computing an approximate solution to the top-k nodes problem since it is hard computationally. We showed the efficacy of this algorithm using a real world coauthorship network from e-print arXiv (www.arxiv.org).

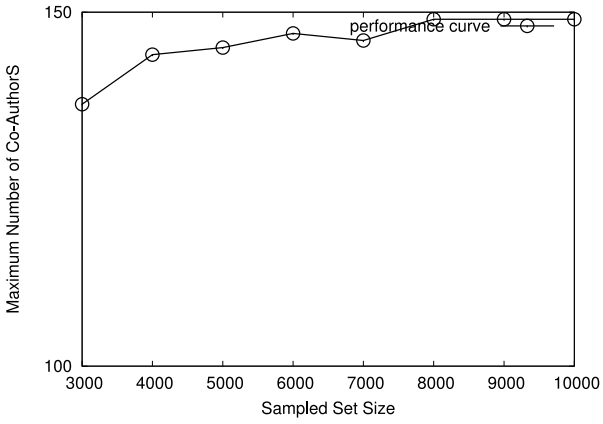

Figure 2: Maximum Number of co-authors for different sampled set sizes when k=15

The proposed approach can also be used even when the underlying objective function is not submodular. This would mean that the approach can be used in a wide variety of related problems.

It would be interesting to study the approximation guarantees provided by *Algorithm 2* more formally. It would also be useful to determine bounds on the size of sampled set (t) of permutations in *Algorithm 2* to get a desired quality of the approximation.

6. REFERENCES

[1] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of ACM SIGKDD*, 2001.

[2] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of ACM SIGKDD*, 2003.

[3] D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *Proceedings of ICALP*, 2005.

[4] R. Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, 1997.

[5] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[6] M. Newman. The structure of scientific collaboration networks. In *Proc. Natl. Acad. Sci.*, 2001.

[7] L. Shapley. *A Value for N-Person Games.* In Kuhn and Tucker, editors, Contributions to the Theory of Games, 1953.

[8] S. Wasserman. *Social Network Analysis.* Cambridge University Press, 1994.