Selection of K-Seeds that give combined Maximum ‘Value’

Viral Team - Sayaji Hande et al (Adobe Research - Bangalore) & R B Bapat (Indian Statistical Institute)

July 22, 2013

Abstract

In a given network, the selection of a set of K nodes that will have the maximum expected spread is a common problem. Be it creating a viral campaign, having check-posts in city to catch criminals or selecting centers in a country for logistic operations. For a given influence matrix, P, and a given a method M of selecting best 'seed', this paper provides a iterative way of forming team of $'K'$ nodes that would provide optimal spread. The research till date has built teams by adding new team member by evaluating 'incremental value. In this paper, we provide 'novel' way of padding all competing members with value of existing team and choosing best among them.

Key Words: Opinion makers, Influence maximization, Election Strategy, Viral Marketing

1 Introduction

Consider a network with N nodes, where each node can influence any other node with some probability. Let $P = ((p_{ij}))$ (possibly asymmetric) define the influence matrix, where p_{ij} indicates the probability that node i influence node j ($p_{ij} \in [0, 1]$). Also, assume that i has the possibility of influencing j only once (as long as $p_{ij} > 0$).

The above formulation arises in many situations, for example, in planning for a product launch on a social network or designing an election campaign. For the success of a campaign on a network, one may want to identify a set of K individuals who will be able to spread a message in a large subset of the network. Similarly, it may be used to identify the most important nodes in a communication network (for example, where to post traffic police in a road network). In other words, given K how does one find that set of K individuals for whom the ‘expected spread’ is maximized.

Each individual i has an opportunity to influence j as long as $p_{ij} > 0$. Consider a situation where $p_{ik} = 0$, then individual i cannot influence k directly. But, if $p_{ij} > 0$ and $p_{jk} > 0$, then i has the possibility of influencing k indirectly through j. Thus for a network of size N, all influences would spread in a maximum of $N - 1$ steps. The ‘expected spread’ of a set of K nodes given a matrix P is the expected number of nodes of the network that are influenced (after $N - 1$ steps).

The idea behind our approach is the following, can we identify the the set of K individuals in an incremental fashion? That is, can we first identify the best subset of size 1. Given this node, we can add the node which has the best incremental spread. This process can be performed iteratively, until a set of size K is reached.

Note that computational complexity of an iterative process is substantially lower. For example, if the population has N nodes and one has to select K seeds, the best subset of size K can be chosen based on $\binom{N}{k}$ comparisons. An iterative process requires only KN operations.

The large part of existing literature either deal with deterministic graphs or convert probabilistic graphs to deterministic and derive solution, see for example, [2] (2003), [5], or references there in. [6] discusses difficulty in computing incremental value add competition of a new node. This paper will provide you easy to use and elegant way of team formation in iterative way. Goyal et al [1] focus of estimation of matrix P based on data.

We try to show that our methods performs close to optimal in many situations.

2 Top Seed Selection - Existing methodologies

Now, lets focus on selection of the top seed. What is the top seed, if one has to select only one individual? Answer to this question is obvious. Let i_0 be a seed. Let X_1 represent # of individuals getting directly influenced by i_0. Similarly X_2 be # of individuals getting indirectly (and not directly), at 2^{nd} level influenced by i_0 through some $j \neq i_0$, and, let X_3 be # of individuals getting influenced at 3^{rd} level and so on. Clearly, for given a matrix P, expected values of $E(X_r)$ for $r = 1, 2, \ldots$ can be computed with ease. This will be described in next section. Top seed is the one with largest value of $\sum_r E(X_r)$.

In other way, if A^r_{ij} is an event that j^{th} individual getting influenced by a seed, say i, at level r^{th} level, all one needs to do is computed, for seed i, $\sum_j \sum_r P(A^r_{ij})$, say, a componant of matrix, $\mathcal{P} = (\mathcal{P}_{ij})$, and then choose i where this sum is maximum.

Even though we are using independence model, *i.e.*, all influence attempts are independent, computing $\mathcal{P}_{ij}$ is algebraically complex. If $P^{(r)}$ represents probabilities of i^{th} infecting j^{th} node exactly at r^{th} level, then

$$\mathcal{P} = \sum_{r=1}^{\infty} P^{(r)}$$

Following is an obvious theorem.

Theorem 2.1 *If S is a seed vector, with 1's and 0's, $\mathbf{1}$ be a vector of 1's and P is influence matrix, then approximated influences of the seed vector S is:*

$$\textit{Total Influence} = f(P, S) \quad = \quad S'\mathcal{P}\mathbf{1}$$

For example, $p^{(2)}_{ij}$, a probability that the j^{th} individual getting influenced by i at secondary stage (alone) is $\sum_l p_{il}p_{lj}(1 - p_{ij})$. Similarly 3^{rd} level, $p^{(3)}_{ij} = \sum_l p^{(2)}_{il} p_{lj}(1 - p_{ij})(1 - p^{(2)}_{ij})$ and so on.

Now, note that, in matrix P, p_{ij}'s are small and $p_{ii} = 1$ for all i, and if p^k_{ij} is an element of natural mutiplication derived matrix P^k, then

$$p^2_{ij} = p_{ij} + \sum_{l \neq j} p_{il}p_{lj} \geq p_{ij} + p^{(2)}_{ij} \geq p_{ij} \tag{1}$$

On similar lines, in general

$$p^r_{ij} \geq p_{ij} + p^{(2)}_{ij} + \cdots + p^{(r)}_{ij} \geq p^{r-1}_{ij} \tag{2}$$

Given above discussions, the $\mathcal{P}$ can be replaced by P^k for some large k for the computations. Under this assumption, following is process of top-seed selection

Selection of Top Seed for Given Matrix P: *If z is Perron vector*[1] *of P, if z_{i_0} is maximum coordinate of z, then, i_0 is a seed that would maximize $s'P^k\mathbf{1}$ for large k.*

This problem is similar to that of selection of top player (or ranking of players) or ranking of web pages. These methodologies have been widely used. See for example, Spizzirri (2011) [7] [2]. Till now, justifications have been provided for deterministic frameworks. In section below we provide justification for the above process for probabilistic networks under discussions.

[1] Eigenvector associate with largest eigenvalue

[2] PageRank: Standing on the shoulders of giants

PerronFrobenius theorem: $http: //en.wikipedia.org/wiki/Perron\%E2\%80\%93Frobenius_theorem$

PageRank: $http: //en.wikipedia.org/wiki/PageRank$

2.1 Application of Perron vector

If A is a square matrix with positive entries then it has a positive eigenvalue, denoted $\rho(A)$, which exceeds any other eigenvalue in modulus. Further it has a positive eigenvector, unique up to a scalar multiple. Thus there exists $x > 0$ such that $Ax = \rho(A)x$. If $B = \frac{1}{\rho(A)}$, then B^k converges to a matrix with columns proportional to x. These results are part of Perron-Frobenius theory. The results hold when A has zeros but is irreducible (which means that for any ij, $p^k_{ij} > 0$ for some k.)

In our problem it is natrural to assume p_{ij}'s to be small, say less than 5 percent. Then $p^{(2)}_{ij}$ in (2) is approximated by $\sum_\ell p_{i\ell}p_{\ell j}$, ignoring higher order terms. Then $\sum_{k=1}^{r} P^{(k)} = P^r$. If $\rho(P)$ is the Perron eigenvalue of P with eigenvector z, then ranking the components of z gives a good indication of the strength of opinion makers. For example if z_i is the maximum component then i is most influential. This is made more precise now.

Think of the initial seed vector as a probability vector s. (If we want a single individual as a seed then the vector has all zeros except a single 1.) For a fixed k, $s'P^k\mathbf{1}$ is the expected number of influenced individuals after exactly k steps. Note that for large k, there will not be much difference between "exactly k" steps and "at most k steps". So our problem is to maximize $s'P^k\mathbf{1}$ over all seed vectors s, where k is fixed and large.

By standard results in Perron-Frobenius theory (such as (2.3) in Keener, [4]), the largest component of the vector $P^k\mathbf{1}$ is i_0 if the i_0-th component of the Perron vector of P is the largest. Let z be the Perron vector of P (which satisfies $Pz = rz$.) If z_{i_0} is the maximum coordinate of z, then the seed vector s_0 which has 1 at place i_0 and zeros elsewhere solves the maximization problem mentioned in the previous paragraph.

So the initial seed can be taken to be i_0 which corresponds to the largest coordinate of the Perron vector of P.

In following sections, we will evaluate, how one can add team members. Or evaluate, after selection of set of 'seeds', what is that 'residual' influence matrix P' one would derive to select another top seed from remaining population or individuals.

3 Process of Creation of K seed set based on method M

Incremental comparisons by adding new member (node) etc has been addressed in literature[3]. Incrementally building 'team' of K seeds has been addressed in literature, Kempe et al [2] (2003). They 1^{st} establish that the 'spread' function is 'submoduler' and then using this they show the 'stepwise' creating of a team provides minimum of $1 - 1/e$ efficiency of best possible K seeds.

This paper too moves in same direction. However, instead of looking at incremental value of next selection, it pads all the competing members with powers of existing member.

Definition 3.1 *A function f is called submoduler, if $S \subset T$ and*

$$f(S \cup x) - f(S) \geq f(T \cup x) - f(T)$$

Definition 3.2 *Let G be a graph with N nodes, P be a influence matrix. Let A be subset of nodes, then, $spread(A)$ is expected # of influences via A nodes.*

It is clear that $spread()$ is a 'Lebesgue measure'. In following we prove that a 'spread' function, that gives expected value of spread for a given subset is 'submoduler'. Kempe et al [2] (2003) did prove that 'spread' function is 'submoduler', following is simpler proof.

Theorem 3.1 *The $spread$ function defined above is 'submoduler'*

[3] Kempe, Kleinberg & Tardos: Influential Nodes in a Diffusion Model for Social Networks

Proof 3.1 *We will work on set theoretic approach and then take expected values of those nodes spread. Please note that if $S \subset T$ and if A is any other subset, then $A \cap S \subset A \cap T$. Hence,*

$$\begin{aligned} S \cup A &= S + S^c A \\ &= S + S^c T A + S^c T^c A \\ \textit{since } S \subset T & \\ S \cup A - S &= S^c T A + T^c A \\ &= S^c T A + (T \cup A - T) \end{aligned}$$

By taking expectations on both sides for the 'spread' by nodes, the result follows.

Hence, as follows from Kempe et al [2] (2003) & [3], any process that maximizes incremental team addition by maximizing incremental spread is at least $100(1 - 1/e)\%$ of overall K team size best.

SH 1/21: Word of caution Following processes provide extending existing team by a team member in best possible manner. However, it is to be noted that, the 'monotonicity' of team formulation may not be possible in some networks. This will be elaborated further in discussions. Enough to say that, if there exist a best team of size, say 10. It does not mean, best team of size 5 would be subset of the same team[4].

Discussions:

Selection: Iterative SH-Method of $K-$Node Selection:

1. To address this issue, we will describe how one can go stepwise 'top seed selection' process by recomputing influence matrix P. Let j_0 be a top seed, selected by method M. Then in matrix $P = ((p_{ij}))$, replace p_{ij_0} by 1 for all i. What this means is, if ANY one of the i is selected as influencer, for this new matrix P, then, she is armed with all influencing powers of i_0 for all subsequent steps. This is true as, she has power of influencing i_0 with probability 1.

2. After selection of 2 or more seeds, replace those respective columns elements with 1's, and select top seed for the reformatted matrix P

3. Execute Step 1 & 2 till you get K seeds

To elaborate further. Lets assume you have a method of selecting the best seed from given population of N individuals and given influence matrix P. Please note that i, jth element of this matrix indicates probability of individual i able to influence individual j. Note that if individual 1 is selected as seed. He or she may directly influence some j or indirectly. P represents the direct probabilities. Indirect probabilities can be computed and are in attached document(s). Here the top seed is the seed, where total # of direct and indirect (secondary, treasury and so on) realized influences are maximum. In absentia of realized influences, we use statistical expected value of influences.

SH 1/21: Elimination On similar lines, given team of K size, we can extend above process of eliminating with least lost and create team of size $K - 1$ from existing team.

The key factor in this formulation is, if there exists a method M, to select the top seed for a given matrix, this incremental 'team creation method' would enable you to create a team of k-seeds, with same complexity of method M. Now, if the method M is most efficient for selecting top seed, this method would provide you with most efficient method to select k-seed which are best.

[4] Ritwik - we need to take pot-shots @ hong-Kong paper as well

As stated earlier, selecting team of k-seeds is substantially different problem that selecting top k ranked seeds. Here we are looking at combined strength of the team, not individual. Hence, one need to maximize the unions. This effectively means, minimizing overlaps.

Now, to understand the solution, how it works. Lets consider individual 1 is identified as top seed at 1st stage. Then, key aspect for selecting next team member, is having effectively arming every individual with all the strength of 1st individual of existing team. To do this, since we know one has been selected this means he is already influenced. There is no harm assuming, he can be influenced by any of the new individual with probability one. Hence, set that column related to individual with all elements equal to 1. What does above do? If you select 3 it is equivalent to selecting 1 & 3, as 3 can influence with probability 1. Hence when 3 & 4 are compared in 2nd step, we know it is comparison of teams $(1, 3)$ & $(1, 4)!!$ This means, by that change in matrix, one would be automatically looking at combined strength in every step. Also, please note that, no method will select individual 1, when she is influence-ble with probability 1 by everyone in population.

SH: 1/21 Following is an important theorems based on above discussions and would lead to substantial decrease in computational efforts. They help only for extending a team by a member or decreasing a team by a member.

Theorem 3.2 $SH-$ **Method** *provides 'Best Incremental Selection* [5]*' if there exist a method* $'M'$ *to select a best seed. Meaning, if* P *is a matrix, and* k *seeds* $j_0, j_1, \ldots, j_k$ *are preselected. Then, if new matrix* $P' = ((p_{ij}))$*, where* $p_{ij_l} = 1$ *for all* $j_0, j_1, \ldots, j_k$ *and computed best seed for this matrix, would provide net maximum influencer to the team.*

Proof 3.2 *The proof of this is obvious.*

Similar to above, following is statement for decreasing a existing team by one member with minimum loss.

Theorem 3.3 $SH-$ **Method** *provides 'Best elimination of a member*[6] *if there exist a method* $'M'$ *to select a best seed. Meaning, if* P *is a matrix, and* k *seeds* $j_0, j_1, \ldots, j_k$ *are preselected. Then, if new matrix* $P' = ((p_{ij}))$*, where* $p_{ij_l} = 1$ *for all, but one of,* $j_0, j_1, \ldots, j_k$ *and computed least incremental seed for this matrix, would provide elimination to the team.*

SH: 1/21 As stated earlier, under some conditions, recursive formation of team may be optimal. Else, adding one team member at a time or so, may not provide an optimal team. For example, if we start with team of size 0 to K, it should be noted that, as we extend selection by way of selection of later members become more restrictive. Hence, to overcome this situation, we provide following *heuristic* process. This way, the 1^{st} member has most dominating position where as last member has least. Hence, following could be a process.

Team Formation Process

1. Start with team size $= 0$, form team based on 'SH-selection' method, till team size reaches K.
2. Use elimination of a least contributing member by 'SH-elimination' method and add a team member by 'selection' method to team of $K - 1$ size. Repeat this process K times.

References

[1] Francesco Bonchi Amit Goyal and Laks V. S. Lakshmanan. Learning inuence probabilities in social networks. *WSDM10, New York City, New York, USA*, February 46, 2010.

[2] Eva Tardos David Kempe, Jon Kleinberg. Maximizing the spread of influence through a social network. *SIGKDD*, 2003.

[5] Assume you have already selected K seeds, looking for $(K + 1)^{th}$

[6] Assume you have already selected K seeds, looking for $(K - 1)$ size team

[3] L Wolsey G. Nemhauser. Interger and combinotorial optimization. *John Wiely*, 1998.

[4] James P. Keener. *SIAM Review*, 35(1):80–93, Mar., 1993.

[5] Thomas W. Valente Raghuram Iyengar, Christophe Van den Bulte. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, March 2011.

[6] Y. Narahari Ramasuri Narayanam. A shapley value based approach to discover inuential nodes in social networks. *Automation Science and Engineering, IEEE Transactions on*, Jan. 2011.

[7] Leo Spizzirri. Justification and application of eigenvector centrality.

4 Appendix

4.1 Application of Perron vector

If A is a square matrix with positive entries then it has a positive eigenvalue, denoted $\rho(A)$, which exceeds any other eigenvalue in modulus. Further it has a positive eigenvector, unique up to a scalar multiple. Thus there exists $x > 0$ such that $Ax = \rho(A)x$. If $B = \frac{1}{\rho(A)}$, then B^k converges to a matrix with columns proportional to x. These results are part of Perron-Frobenius theory. The results hold when A has zeros but is irreducible (which means that for any ij, $p_{ij}^k > 0$ for some k.)

In our problem it is natrural to assume p_{ij}'s to be small, say less than 5 percent. Then $p_{ij}^{(2)}$ in (2) is approximated by $\sum_\ell p_{i\ell}p_{\ell j}$, ignoring higher order terms. Then $P^{(r)} = P^r$. If $\rho(P)$ is the Perron eigenvalue of P with eigenvector z, then ranking the components of z gives a good indication of the strength of opinion makers. For example if z_i is the maximum component then i is most influential. For a set of k seeds we may choose the top k components of z.

If this approach makes sense then one can explore how matrix P may be partitioned and opinion makers in different territories may be chosen.

4.2 Incremental Addition of Seeds may not lead to Optimal Subset - A Counter example

Consider a graph with $2N + 1$ nodes with the influence matrix given by the matrix below.

$$P = \left(\begin{array}{ccc} 1 & \mathbf{p}_1' & \mathbf{p}_1' \\ \mathbf{0} & \boxed{\begin{array}{cccc} 1 & p_2 & \cdots & p_2 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{array}}_{N\times N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boxed{\begin{array}{cccc} 1 & p_3 & \cdots & p_3 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{array}}_{N\times N} \end{array}\right)_{(2N+1)\times(2N+1)}$$

Where $\mathbf{p}_1$ is the vector $(p_1, \cdots, p_1)'$. Such a network has three groups of individuals. The first individual (1) has an influence over all nodes (a global influencer). The next two sets of nodes (given by $\{2, \cdots, N+1\}$ and $\{N+2, \cdots, 2N+1\}$), forms two groups with only one node that has an influence. Without loss of generality assume that $p_2 = p_3 = p$.

Define $E(S_\Sigma)$ as the expected spread of the set $\Sigma \subseteq \{1, 2, \cdots, 2N+1\}$. Note that,

$$\begin{aligned} E(S_{\{1\}}) &= 1 + 2Np_1 + \Sigma_{i=3}^{N+1}(1-p_1)p_1p_2 + \Sigma_{i=N+3}^{2N+1}(1-p_1)p_1p_3 \\ &= 1 + 2Np_1 + (N-1)(1-p_1)p_1p_2 + (N-1)(1-p_1)p_1p_3 \\ &= 1 + 2Np_1 + 2(N-1)(1-p_1)p_1p \qquad \text{if } p_2 = p_3. \end{aligned}$$

Further,

$$\begin{aligned} E(S_{\{1,2\}}) &= 2 + \Sigma_{i=3}^{N+1}(p_1 + p_2 - p_1p_2) + Np_1 + (N-1)(1-p_1)p_1p_3 \\ &= 2 + (N-1)(p_1 + p - p_1p) + Np_1 + (N-1)(1-p_1)p_1p \qquad \text{if } p_2 = p_3. \end{aligned}$$

and

$$E(S_{\{1,N+2\}}) = 2 + 2(N-1)p$$

Then, can we construct a situation such that, for all i, but for some i, j and k.
It is possible to show that for some choices of N, p_1 and p, we have a situation that

$$E(S_{\{1\}}) > E(S_{\{2\}})$$

but

$$E(S_{\{1,2\}}) < E(S_{\{2,3\}}).$$

One choice that leads to this situation is when $N = 1000$, $p_1 = 0.04$, $p = 0.08$. Then we have the following,

$$\begin{aligned} E(S_{\{1\}}) &= 87.14, \\ E(S_{\{2\}}) &= 80.92, \\ E(S_{\{1,2\}}) &= 161.75, \\ E(S_{\{2,3\}}) &= 161.84. \end{aligned}$$

Thus, for such an influence matrix, starting with the most influential node (1) and incrementally adding members to this seed set leads to a size 2 set which is not optimal.

4.3 Wrong Proof

Theorem 4.1 *$SH-$ **Method** provides 'Best K seed Selection' if there exist a method $'M'$ to select a best seed.*

Proof 4.1 *The discussions prior to theorem are logically complete.*
However, in following, this proof provides clear steps. We will prove for case of $K = 2$, other cases follows.

1. *Lets assume there is a graph G, a relationship / influence matrix P and best two seeds (S_1, S_2).*
2. *It is obvious that if one desires to select $(K+1)^{th}$ seed given existing 'top selection' method M, above process would give 'best' $(K+1)^{th}$ addtional 'seed'.*
3. *Assume there are two seeds set, S_1 and S_2, say, (S_1, S_2) is best. Assuming, without loss of generality, all seeds are 'unique'*[7]
4. *Do $SH-$Method on S_1 alone. Lets assume you get a set (S_1, S_2'). Since, (S_1, S_2) is best seed set, $S_2' == S_2$, due to uniqueness assumption.*
5. *Similarly, do $SH-$Method on S_2 alone. This will give you $S_1' == S_1$.*
6. *Now we know $E^8(S_1 + S_2) == E(S_1' + S_2) == E(S_1 + S_2')$*
7. *Now create a graph $G_1 + G_2$, with $G_1 = G_2 = G$, basically two disconnected copies of G. Place (S_1, S_2) in G_1. Now, use $SH-$method. You will end up with (S_1'', S_2''). This gives you, $E(S_1 + S_2) == E(S_1'' + S_2'')$*

[7] This is easy to achieve by adding ϵ's to P_{ij}'s and letting it tend to zero

[8] This is expected value, and addition is not a simple addition but, combined value of seeds

4.4 Perron-Frobenius Theorem and Related Results

In following we state the *Perron-Frobenius* theorem.

Theorem 4.2 *If M is an $n \times n$ nonnegative primitive matrix, then there is a largest eigenvalue λ_0 such that*

1. *λ_0 is positive*
2. *λ_0 has unique (up to a constant) eigenvector vector v_1, which may be taken to have all positive entries*
3. *λ_0 is non-degenerate*
4. *$\lambda_0 > |\lambda|$ for any eigenvalue $\lambda \neq \lambda_0$*

Given above *Perron-Frobenius* theorem and discussions, only for a given matrix P, we can derive, how to select a top seed.

Theorem 4.3 *If P is an $n \times n$ nonnegative influence matrix, if S is a seed vector of 0's and 1's then $\max\limits_{\mathbf{1}'S=1} f(P,S)$ is achieved if*

1. *λ_0 is largest eigenvalue*
2. *v_1 is eigenvector vector associated with λ_0*
3. *S is vector where there is 1 at index associated with largest value in components of v_1 and rest are 0's*

Proof 4.2 *Let λ_i's and v_i be respective eigenvalues and eigenvectors. Then for any vector x. Since v_i's are orthogonal, we will have α_i's such that, $x = \sum \alpha_i v_i$and*

$$\begin{aligned} P^k x &= \sum \lambda_i^k \alpha_i v_i \\ \frac{P^k x}{\lambda_0^k} &= \alpha_0 v_0 + \sum_{i \geq 1} \frac{\lambda_i^k \alpha_i vi}{\lambda_0^k} \end{aligned}$$

Now, letting $k \to \infty$, $\frac{P^k x}{\lambda_0^k} \to \alpha_0 v_0$ as $\lambda_0 > \lambda_i$ for all $i \neq 0$.
This completes the proof.

Above results, for deterministic graphs, where P has only 0 or 1's as entries, is called 'eigenvalue' centrality.
Please note that, we have to select k nodes. In this, we established, for given a matrix P, how to establish a top node.