Eigenvector Centrality for Decomposable Markov Chains

Kristen Mecadon*, Megan Schneck°, Annalinda Arroyo*,
Karl Heinz Hoffman†, Jim Nulton*, Peter Salamon*

July 26, 2007

* SDSU
° Baylor University
† Chemnitz

Abstract

Eigenvector importance ranking allows us to define ranks for states of a random walk. If the states are disconnected we have a decomposable walk. The shadow graph can be used to make connections and compare the ranks in different components. The biological motivation for this technique is the graph derived from the distance matrix on phage proteins with the shadow graph linking all proteins in the same phage.

1 Introduction

Eigenvector based importance ranking [1] is used in many applications such as marketing, influence flow, belief networks, and web searches. The items to be ranked are nodes in a network. The importance of the j^{th} node is measured by the j^{th} entry in the principal eigenvector of the matrix describing the network. For influence ranking, the ij^{th} entry of the matrix describing the network measures the extent of influence that node i exerts on node j. In this paper, we present a method for combining importance ranks on disjoint sets. The rankings of the nodes in each disjoint set remain invariant. Our technique connects the disjoint sets and re-ranks the nodes relative to all other nodes. This re-ranking is achieved by using a relation G on the union of the disjoint sets. We refer to the effect of G as a *shadow structure*.

Consider m disjoint random walks on sets $S_i, i = 1, \ldots, m$ with transition matrices $M_i, i = 1, \ldots, m$. The dominant eigenvectors, V_i, of these disconnected random walks define the rankings in each set. Formally, we may always write a transition matrix

$$M = \begin{pmatrix} M_1 & 0 & \ldots & 0 \\ 0 & M_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & M_m \end{pmatrix}$$

on the union $S = \cup_i S_i$. Now consider a relation G on S. The matrix G connects the disjoint Markov chains by making small perturbations on the disjoint sets S_i. This connection allows us to re-rank the nodes in S. As the perturbation approaches zero, a "shadow" of these connections remains. We find a shadow vector which is a linear combination of the stationary eigenvectors of the disjoint sets. We define V_i as the unique stationary distribution of the set M_i. The linear combination gives new ranks of the nodes relative to all other nodes. Our motivation comes from a bioinformatics problem: the desire to rank proteins in bacteriophage, i.e., in viruses that predate on bacteria. Transition probabilities between proteins can be derived from standard bioinformatic methods based on similarity or homology. The Markov chain, associated with these transition probabilities, is decomposable since the proteins fall into different, mutually non-homologous families. The relation G is used to connect proteins that are present in the same bacteriophage.

2 The Formal Problem

We begin with definitions that introduce the notation needed in the arguments below.

Definition 1. *Given a matrix X, let X_0 denote the diagonal matrix whose j^{th} entry is the sum of the entries in the j^{th} column of X. Note that the operator $(\ )_0$ is linear.*

Definition 2. *Given non-negative $n \times n$ matrices M and G, and $\epsilon > 0$, define*

$$M_\epsilon = (M + \epsilon G)(M + \epsilon G)_0^{-1}. \tag{1}$$

For our problem, M corresponds to a block diagonal transition probability matrix of a decomposable Markov chain as shown in (1). We assume that each of the nonzero blocks is a regular chain [5]. The matrix G perturbs the decomposable Markov chain and relates the disjoint components. We assume that M_ϵ is indecomposable for $\epsilon > 0$. By using G to merge the distinct Markov processes, we rank the importance of all n nodes in the connected chain using the unique stationary distribution of M_ϵ. In the limit, as ϵ goes to zero, the stationary distribution of any component S_i remains unchanged.

Definition 3. *A* shadow structure *is a pair (M, G), where M is the matrix of a decomposable Markov chain which is regular on each component and G is a non-negative matrix such that M_ϵ is regular for $\epsilon > 0$.*

Definition 4. *The* shadow vector *V for the shadow structure (M, G) is the limit of the unique stationary distribution of M_ϵ, as $\epsilon \to 0$.*

3 The Proof

To calculate and interpret the shadow vector V we will use the technique of lumping in Markov chains [5]. The techinique lumps the nodes in a chain by adding incoming probabilities and averaging outgoing probabilities. Each probability is re-normalized by weights corresponding to the stationary distribution. In terms of matrices, this is accomplished by

$$\text{Lump}(M) = CMD \tag{2}$$

using a collect matrix C that adds the rows and a distribute matrix D that averages the columns. For a shadow structure (M, G), we are interested in lumping all the states in the i^{th} component of the decomposable chain M.

Definition 5. *Define the collect matrix C of (M, G) as the $m \times n$ matrix of ones and zeros with the ij^{th} entry a one if the j^{th} column of M contains an entry of the i^{th} block of M and a zero otherwise.*

Definition 6. *Define the distribute matrix D of (M, G) as the $n \times m$ matrix whose j^{th} column is the normalized eigenvector of M whose support is the j^{th} block of M.*

Notice that the j^{th} column of D is V_j padded with zeros in other blocks. The following lemma asserts that the linear operators Lump and $(\ )_0$ commute. We will need this fact in order to prove Theorem 3. Although C and D are defined for our specific case, the following lemma is true for any C and D.

Theorem 1. *For any collect and distribute matrices C and D and an $n \times n$ matrix G, $(CGD)_0 = CG_0D$.*

Proof. Let 1_p denote the $1 \times p$ vector of ones. Notice that, for any matrix $S_{p\times q}$, 1_pS is a row vector whose j^{th} entry is the j^{th} column sum of S.

Now, observe that

$$1_nG = 1_nG_0. \tag{3}$$

Also, notice that

$$1_mC = 1_n, \tag{4}$$

since the columns of the $m \times n$ collect matrix sum to one.

Since CG_0D is a diagonal matrix, then, in order to show that $(CGD)_0 = CG_0D$, it suffices to show that CDG and CG_0D have the same column sums. Thus, it suffices to show that $1_mCGD = 1_mCG_0D$. Observe that both CDG and CG_0D are $m \times m$ matrices. Then, using (3) and (4),

$$1_mCGD = 1_nGD = 1_nG_0D = 1_mCG_0D. \tag{5}$$

Therefore, CDG and CG_0D have the same column sums. So, $(CGD)_0 = CG_0D$. □

We hope to use this theorem to prove our main result. Next, we formally define a shadow vector as a stationary eigenvector of the Markov chain with n states.

Definition 7. *A vector V is a shadow vector for M with respect to G iff $V = \lim_{\epsilon \to 0} \mu_\epsilon$ where $M_\epsilon \mu_\epsilon = \mu_\epsilon$.*

We use this formal definition to prove Theorem 2, which functions as the main result of this section.

3.1 Example

In order to motivate our main result, consider the application of our method to the two block case. We illustrate a method for finding v which is the weight on the stationary distribution. We begin by lumping the connected Markov chain $[M + \epsilon G]$. Recall that M,G are $n \times n$ matrices, C is the collect matrix and D is the distribute matrix. In this example, since $S = S_1 \cup S_2$, C is a $2 \times n$ matrix and D is an $n \times 2$ matrix. Notice,

$$
\begin{aligned}
C[M + \epsilon G]D &= C\left[\begin{pmatrix} M_{11} & 0_{d_1 \times d_2} \\ 0_{d_2 \times d_1} & M_{22} \end{pmatrix} + \epsilon \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}\right] D \qquad (6) \\
&= C\begin{pmatrix} M_{11} + \epsilon G_{11} & \epsilon G_{12} \\ \epsilon G_{21} & M_{22} + \epsilon G_{22} \end{pmatrix} D \qquad (7) \\
&= \begin{pmatrix} CM_{11}D + \epsilon CG_{11}D & \epsilon CG_{12}D \\ \epsilon CG_{21}D & CM_{22}D + \epsilon CG_{22}D \end{pmatrix} \qquad (8) \\
&= \begin{pmatrix} 1 + \epsilon P_{11} & \epsilon P_{12} \\ \epsilon P_{21} & 1 + \epsilon P_{22} \end{pmatrix} \qquad (9)
\end{aligned}
$$

It is important to note that after the above step, we re-normalize the connected set so that the column sums of the matrix add to one.

This example illustrates that the entries of v are the off-diagonal entries of $P = CGD$. Thus, v satisfies the detailed balance condition, which the 2×2 block case always satisfies. Observe that $P_{21}v_1 = P_{12}v_2$, for $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$, and therefore $\dfrac{v_1}{v_2} = \dfrac{P_{12}}{P_{21}}$.

Now we state our main result, Theorem 2.

Theorem 2. *Given a nullvector v of $C(G - G_0)D$, let $V = Dv$, where D denotes the distribute matrix of M. Then, V is a shadow vector for (M, G).*

This theorem is true, but does not yet have a solid proof.

These results provide a method for finding a shadow vector of a decomposable Markov chain M. Now, we consider the case where $M = I$. Then, the connected transition probability matrix is $M_\epsilon = (I + \epsilon A)(I + \epsilon A)_0^{-1}$, where A is any perturbation matrix. The following theorem characterizes the shadow vector of the lumped Markov chain, for the case where $M = I$.

Theorem 3. *If $(A - A_0)v = 0$, then v is a shadow vector for (I, A).*

Proof. Suppose v is a nullvector for $(A - A_0)$. Let $\mu_\epsilon = (I - \epsilon A)_0 v$, and note that $v = \lim_{\epsilon \to 0} \mu_\epsilon$. It remains to show that $M_\epsilon \mu_\epsilon = \mu_\epsilon$. Calculating, we have

$$M_\epsilon \mu_\epsilon = (I + \epsilon A)(I + \epsilon A)_0^{-1}(I + \epsilon A)_0 v = (I + \epsilon A)v. \tag{10}$$

Moreover, note that

$$[(I + \epsilon A) - (I + \epsilon A)_0]v = \epsilon(A - A_0)v = 0. \tag{11}$$

Therefore,

$$(I + \epsilon A)v = (I + \epsilon A)_0 v = \mu_\epsilon. \tag{12}$$

Combining (10) and (12) completes the proof. □

The following theorem characterizes the relationship between the shadow vector V for the connected chain and the shadow vector v for the lumped connected chain.

Theorem 4. *If $V = Dv$ is a shadow vector of (M, G), where D is the distribute matrix of M, then v is a shadow vector of (I, A).*

The proof for Theorem 4 is in progress. We hope to use the proof for Theorem 3 to prove the above statement. However, we can speak about the elements of the shadow vector v for the lumped Markov chain.

Theorem 5. *The components of v can be taken to be the diagonal minors of $(A - A_0)$.*

Proof. Define $B = A - A_0$. Recall, from Theorem 4, that if $Bv = \vec{0}$, then v is a shadow vector. In order to show that the components of v can be the diagonal minors of B, it suffices to show that, if v_i denotes the *ith* diagonal minor of B, $Bv = \vec{0}$. Now, observe that B is a singular matrix, since the columns of B sum to zero.

Case 1. *Suppose that* $rank(B) \leq m-2$. *Since* B *is a singular matrix,* $det(B) = 0$. *Then, using the formula*

$$adj(B)B = Badj(B) = det(B), \tag{13}$$

we have $adj(B) = 0_{m\times m}$, *where*

$$adj(B) = [(-1)^{(i+j)} detB(i|j)]^T. \tag{14}$$

So, $\vec{0} = \begin{bmatrix} \beta_{11} \\ \beta_{22} \\ \vdots \\ \beta_{mm} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$, *where* β_{ii} *is the* i^{th} *diagonal entry of* $adj(B)$, *and* $Bv = \vec{0}$.

Case 2. *Suppose that* $rank(B) = m-1$. *Since* B *is singular,* $det(B) = 0$, $adj(B)B = Badj(B) = 0_{m\times m}$, *using* (13). *Now, let* $r = row_i(adj(B))$ *and* $c = column_j(adj(B))$. *Then,*

$$rB = 0_{1\times m} \tag{15}$$

and

$$Bc = 0_{m\times 1}. \tag{16}$$

Furthermore, notice that, for a $1 \times m$ *vector* e *of ones,*

$$eB = 0_{1\times m} \tag{17}$$

because the columns of B *sum to zero. Now, using equations* (13) *and* (15), *and the fact that* $rank(B) = m-1$ *iff* $rank(adj(B)) = 1$, $r = \alpha e$, *for some* $\alpha \in \mathbb{R}$. *Each row is a scalar multiple of the vector* e. *Thus, the entries in a row are the same. So, the diagonal entries of* B *equal the column entries for any column. Then,* $B \begin{bmatrix} \beta_{11} \\ \beta_{22} \\ \vdots \\ \beta_{mm} \end{bmatrix} = Bc = 0_{m\times 1}$. *Therefore, the diagonal minors can be taken to be the components of* v.

□

4 Discussion

In section 3, we state that, for the linked Markov chain $M_\epsilon = (M + \epsilon G)(M + \epsilon G)_0^{-1}$, $V = Dv$ is a shadow vector for (M, G). Also, we hope to prove that if V is a shadow vector for (M, G), then v is a shadow vector for A when $M = I$.

When considering these theorems the question arises 'How does one comprehend the role of the vector v?' According to Theorem 5, v contains the diagonal minors of $(A - A_0)$. This theorem, however, fails to give an interpretation of the role of v. The following two analogies provide a conceptual understanding of the vector v.

Suppose that two families contain seven people. Three people belong to Family A and four people belong to Family B. Each family measures the height of its members and ranks them accordingly. The tallest individual receives the largest rank and the shortest person receives the smallest rank. However, Family A measures height with a meter stick and Family B measures with a yard stick. So, Family A ranks its members after finding heights in centimeters while Family B ranks its members after finding heights in inches. Now, suppose that we desire to rank all seven people, using only the set of heights of each family. Before we assign rankings, we must correctly scale the units of one set of measurements to match the units of the other set. Then, we can compare the numbers and rank accordingly.

This example illustrates the functioning of the vector v, which scales the measurements in order to compare ranks across non-homologous sets. We obtain the importance rankings for the proteins in a certain set by finding the stationary distribution for the Markov chain within that set. The entries in this distribution correspond with the measurements from one family. However, in order to find the importance rankings for nodes in one set relative to nodes in the $m - 1$ other sets, we re-scale the rankings of the $m - 1$ other sets. In terms of the preceding example, we require that the units of measurement agree for both families. Then, v is a vector that scales the rankings from the m different sets. Thus, V is a linear combination of the stationary distribution of the decomposable Markov chain, and v functions as the weights that enable the ranking of all proteins; these scaling factors are the diagonal minors of $(A - A_0)$.

As a second example of the function of v, suppose that fifty people walk between three department stores. They also have the ability to walk between the various sections, for instance men's clothing and women's clothing, within

a store. At each step of time, the people can transition to another section within a department store or to another department store. Lets suppose for this example the people shop for 24 hours. The people move from one store to another store and from one section of the department store to another section with certain probabilities. This random walk between the three stores, (or states) forms a Markov chain. Also, the random walks between the sections in each store form Markov chains.

In this context, the shadow vector ranks the importance of a section in a store relative to all sections of the stores. The vector v is a stationary distribution for the three stores. The elements of v are probability weights that maintain an equilibrium amount of flow in and out of each store. A weight on a store is the fraction of the day that each person spends in that store.

This example illustrates the function of v in our application of Markov processes to protein mutation. A protein set corresponds with a store and proteins correspond with sections. Mutation from one protein to another protein, within a non-homologous family, functions as a Markov chain. Also, mutation from one set to another set is a Markov chain.

Thus, using these two examples, we interpret the function of v as a scaling factor that enables the ranking of the different families of proteins. This scaling factor, or weight, contains the diagonal minors of $(A - A_0)$.

5 Conclusion

This paper presents a method for ranking the n nodes of M, a decomposable Markov chain, with m disjoint Markov chains as blocks along the diagonal of M. We implement a shadow structure in order to link the disjoint Markov chains. The effect of the shadow structure is to merge the disjoint Markov processes. Then, we can find the shadow vector, the unique stationary distribution, of the connected probability transition matrix, M_ϵ. We then use the technique of lumping; we lump the states in the i^{th} component of M_ϵ. Our main result, Theorem 2, states that $V = Dv$ is the shadow vector for M_ϵ, where D is the distribute matrix of M. We hope to prove that if $V = Dv$ is the shadow vector for (M,G), then v is the shadow vector for (I, A), the case where M=I. This theorem states that if V is the shadow vector for the connected chain, before lumping, then v is the shadow vector for the connected chain, after lumping. Furthermore, we prove in Theorem 5 that

the vector v contains the diagonal minors of $(A - A_0)$, for A the perturbation matrix of the lumped chain. We interpret v in two ways. This vector serves as a scaling factor that enables the ranking of the elements of the disjoint chains. Also, v contains the weights that equalize the flow among the lumped components.

References

[1] Stephen Borgatti, 2005 "Centrality and Network Flow♣"

[2] Bonacich, 1972 "Factoring and Weighting Approaches to Status Scores and Clique Identification." *Journal of Mathematical Sociology* 2:113-120

[3] Bonacich, 1987 "Power and Centrality: a Family of Measures." *American Journal of Sociology* 92:1170-1182

[4] Bonacich, 1991 "Simultaneous Group and Individual Centralities." *Social Networks* 13:155-68

[5] Kemeny, John G. Snell, J. Laurie 1960.*Finite Markov Chains*. New York.