PERMUTING MARKOV CHAINS TO NEARLY COMPLETELY DECOMPOSABLE FORM

TUĞRUL DAYAR*

Abstract. This paper highlights an algorithm that computes, if possible, a nearly completely decomposable (NCD) partitioning for a given Markov chain using a specified decomposability parameter. The algorithm is motivated by search for connected components (CCs) of an undirected graph. The nestedness of the NCD partitionings for increasing values of the decomposability parameter is demonstrated on the Courtois matrix. The relation among the degree of coupling, the smallest eigenvalue of the coupling matrix and the magnitude of the subdominant eigenvalue of the block Gauss-Seidel (BGS) iteration matrix induced by the underlying NCD partitionings is investigated on the same matrix. Experimental results that appear elsewhere show that the partitioning algorithm may be used successfully in two-level iterative solvers such as block successive over-relaxation (BSOR) and iterative aggregation-disaggregation (IAD).

Key Words. Markov chains, near complete decomposability, degree of coupling, orderings, block SOR, iterative aggregation-disaggregation

AMS(MOS) subject classification. 60J10, 65U05, 65F30, 65F10

1. Introduction. Nearly completely decomposable (NCD) Markov chains [3], [8], [11] are irreducible stochastic matrices that can be symmetrically permuted to a block form as in

$$P_{n\times n} = \begin{array}{c} \begin{array}{cccc} n_1 & n_2 & \cdots & n_N \end{array} \\ \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{21} & P_{22} & \cdots & P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{N1} & P_{N2} & \cdots & P_{NN} \end{pmatrix} \end{array} \begin{array}{c} n_1 \\ n_2 \\ \vdots \\ n_N \end{array}, \tag{1}$$

where the nonzero elements of the off-diagonal blocks are small compared with those of the diagonal blocks [11, p. 286]. Let $P = \mathrm{diag}(P_{11}, P_{22}, \ldots, P_{NN}) + E$. The diagonal blocks P_{ii} are square, of order n_i, with $n = \sum_{i=1}^{N} n_i$. The quantity $\|E\|_\infty$ is referred to as the degree of coupling and is taken to be a measure of the decomposability of P. When the chain is NCD, it has eigenvalues close to 1, and the poor separation of the unit eigenvalue implies a slow rate of convergence for standard matrix iterative methods [4, p. 290]. The smaller $\|E\|_\infty$ is, the more ill-conditioned P becomes [8, p. 258]. On the other hand, if P were reducible, we would decompose the chain into its irreducible (i.e., isolated) and transient subclasses of states as in equation (1.20) of [11, p. 26] and continue our analysis on the irreducible subclasses. If $\|E\|_\infty$ were zero, then P would be completely decomposable.

Such matrices arise in queuing network analysis, large-scale economic modeling, and computer systems performance evaluation. The long-run measures of interest for these systems may be obtained from the long-run distribution of state probabilities by solving a homogeneous system of linear equations with a singular coefficient matrix under a normalization constraint [11, p. 16].

An algorithm (see [12, Section 3] and [11, Section 6.3.5]) that is currently being used for finding an ordering of states as in equation (1) searches for the strongly connected components (SCCs) of the directed graph (digraph) [6, p. 2] associated with the matrix obtained by zeroing the elements of P that are less than a user specified decomposability parameter ϵ, a real number between 0 and 1. The partitioning of such a graph into its SCCs is unique (see [6, pp. 113–122]).

* Department of Computer Engineering and Information Science, Bilkent University, 06533 Bilkent, Ankara, Turkey (tugrul@cs.bilkent.edu.tr).

The subset(s) of states output by the SCC search algorithm are identified as forming the NCD blocks P_{ii}. As we show in the next section, this algorithm may fail to produce a correct NCD partitioning of the state space, and therefore needs reconsideration.

Among other existing partitioning algorithms that take into account the values of the nonzero elements of the underlying matrix, the one proposed in a different context by Sezer and Šiljak [9, 10] which is motivated by search for connected components (CCs) of an undirected graph may be used to compute NCD partitionings of P. Another algorithm, version 1 of the Threshold Parameterized Block Ordering Algorithm (TPABLO, see Criterion 1 in [2]), has a set of five input parameters, and albeit powerful, is not easy to fine-tune.

In the next section, we show that the SCC search algorithm of the MARkov Chain Analyzer (MARCA) software package (see [11, p. 502]) fails using the 8×8 Courtois matrix. In the third section, we present the algorithm due to Sezer and Šiljak in a Markov chain setting. In the fourth section, we take a close look at the NCD partitionings the CC search algorithm computes for the Courtois matrix and point out the relation among the degree of coupling, the smallest eigenvalue of the coupling matrix and the magnitude of the subdominant eigenvalue of the block Gauss-Seidel (BGS) iteration matrix induced by the underlying NCD partitionings. In the fifth section, we conclude with some observations.

2. A counter-example. Consider the 8×8 NCD matrix [3]

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left(\begin{array}{cccccccc} 0.85 & 0 & 0.149 & 0.0009 & 0 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.65 & 0.249 & 0 & 0.0009 & 0.00005 & 0 & 0.00005 \\ 0.1 & 0.8 & 0.0996 & 0.0003 & 0 & 0 & 0.0001 & 0 \\ 0 & 0.0004 & 0 & 0.7 & 0.2995 & 0 & 0.0001 & 0 \\ 0.0005 & 0 & 0.0004 & 0.399 & 0.6 & 0.0001 & 0 & 0 \\ 0 & 0.00005 & 0 & 0 & 0.00005 & 0.6 & 0.2499 & 0.15 \\ 0.00003 & 0 & 0.00003 & 0.00004 & 0 & 0.1 & 0.8 & 0.0999 \\ 0 & 0.00005 & 0 & 0 & 0.00005 & 0.1999 & 0.25 & 0.55 \end{array} \right) \end{array}$$

which has a degree of coupling equal to 0.001 for the state space partitioning $\{1,2,3\}$, $\{4,5\}$, $\{6,7,8\}$. Now, let $\epsilon = 0.125$. If we zero out the elements in P that are less than ϵ, we obtain a matrix with the following nonzero structure:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left(\begin{array}{cccccccc} \mathrm{X} & 0 & \mathrm{X} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{X} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & \mathrm{X} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{X} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{X} & \mathrm{X} & \mathrm{X} \end{array} \right) \end{array},$$

where an X denotes a nonzero. Conducting an SCC search on the digraph associated with this new matrix yields the partitioning $\{1\},\{2,3\},\{4,5\},\{6,8\},\{7\}$. However, it is not true that the submatrix associated with states $\{6,8\}$ is an NCD block, and hence the symmetric permutation suggested by this state space partitioning is NCD with decomposability parameter 0.125. Both states 6 and 8 have transitions to state 7 with probability greater than 0.125. This example shows that it is not possible to identify SCCs as NCD blocks.

3. The algorithm. The problem with the SCC search algorithm is that it misclassifies subsets {1} and {6,8} as forming NCD blocks. One solution would be to group state 1 with the subset {2,3} to which it has a transition with probability 0.149, and to group states 6,7,8 into a separate subset so that they do not have any transitions with probability larger than or equal to 0.125 to outside states. Thus, we end up with the partitioning {1,2,3}, {4,5}, {6,7,8}. Hence, for the particular value of ϵ under consideration, an NCD partitioning of the Courtois matrix exists with $\|E\|_\infty = 0.001$.

The CC search algorithm due to Sezer and Šiljak provides a simpler way to obtain the same NCD partitioning. First, construct an undirected graph whose vertices are the states of P by introducing an edge between vertices i and j if $p_{ij} \geq \epsilon$ or $p_{ji} \geq \epsilon$, and then identify its CCs (see [9, p. 322]). Since Markov chains that arise in real-life applications are mostly large and sparse, we assume that the input to the algorithm is the matrix P stored in Compact Sparse Row (CSR) Harwell-Boeing format, which requires three arrays: one real and one integer of size nz (i.e., number of nonzeros in P), and one integer of size $n+1$ [11, pp. 80–81]. A simple implementation is then provided by the following:

Algorithm. Finding an NCD form corresponding to a decomposability parameter $0 < \epsilon < 1$ of a Markov chain P:

Step 1. Make one pass over P and P^T simultaneously and form the symmetric boolean matrix A in which $a_{i,j}$ is set to true if $p_{i,j} \geq \epsilon$ or $p_{j,i} \geq \epsilon$. The elements of A that are not set to true are considered false.

Step 2. Search for the CCs of the undirected graph associated with the true elements in A. Each CC is a subset of the NCD partitioning.

One can declare P as non-NCD for the chosen ϵ if all n states end up in the same (and only) subset after step 2. Otherwise, P should be declared NCD.

Note that there is no need to store A as a two-dimensional array. It suffices to store pointers to the beginning of each row in the CSR format (requiring $n+1$ integers) and the column indices of the true elements in A (requiring at most $2nz$ integers), effectively an adjacency list structure. It is possible to do this using a temporary boolean array of length n while scanning the rows of P and P^T, and marking its appropriate entries true according to the criterion in Step 1. When the row of interest is processed, this boolean array of length n may be scanned from left to right and the column indices of its true entries stored in linear storage. Then the boolean array should be cleared for processing the next row of P and P^T. The transpose of P should be held separately in CSR format during Step 1 since P and P^T will be scanned simultaneously in a row-by-row manner. Otherwise, it is necessary to process P stored in CSR format both row-by-row and column-by-column, which is undesirable. Additional space used by the CC search in Step 2 is linear in n [1, p. 181]; hence, the space requirement of the NCD partitioning algorithm is $O(n+nz)$ integers and $O(nz)$ reals.

The time complexity of the transposition of P stored in CSR format in Step 1 is negligible since it does not involve any comparisons. The same step involves $O(nz)$ floating-point comparisons for the test. The time complexity of the CC search in Step 2 is $O(n+nz)$ [1, pp. 180–181]. Therefore, the time complexity of the NCD partitioning algorithm is $O(n+nz)$.

Consider, for instance, the case where P is a normwise very small perturbation of the identity matrix, I. For a myriad of values for ϵ, the output of the algorithm will be the partitioning in which each state forms an NCD block of its own. The result is expected; such matrices are called coupling matrices [8, p. 249] and they arise when NCD Markov chains are aggregated based on an NCD partitioning.

For the Courtois matrix, A is given by

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left(\begin{array}{cccccccc} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array} \right) \end{array},$$

where a 1 indicates a true value. The output of step 2 is the partitioning $\{1,2,3\},\{4,5\},\{6,7,8\}$. We should remark that TPABLO version 1 [2] with the set of parameters $\alpha \in \{0.0, 0.5, 1.0\}, \beta \in \{0.0, 0.5, 1.0\}, \gamma = 0.125, minbs = 1, maxbs = 8$ on the Courtois matrix gives the partitioning {1},{2},{3},{4},{5},{6},{7},{8} in all nine cases. Here, α is the density parameter for diagonal blocks, β is the connectivity parameter among blocks, γ has the same meaning as ϵ, *minbs* and *maxbs* are respectively the desirable minimum and maximum order of diagonal blocks.

The NCD partitioning obtained by the CC search algorithm may be used in two-level iterative solvers for Markov chains such as block successive over-relaxation (BSOR) [11, Section 3.3] and iterative aggregation-disaggregation (IAD) [11, Section 6.3]. Since the partitioning algorithm has linear time complexity in the problem size and does not involve any floating-point operations, the overhead associated with using it as a preprocessing step can be considered negligible.

4. Relationship between degree of coupling and asymptotic convergence rate. Consider the following homogeneous system of linear equations with a normalization constraint

$$\pi(P - I) = 0, \quad \|\pi\|_1 = 1,$$

for the purpose of computing the unknown $(1 \times n)$ stationary vector π.

Now let the block Gauss-Seidel (BGS) splitting of the coefficient matrix $A = P - I$ corresponding to the block form in equation (1) be given by

$$A = L - (D - U), \tag{2}$$

where L, $-D$, U represent respectively strictly block lower-triangular, block diagonal, strictly block upper-triangular parts of A. For the splitting in equation (2) the BGS iteration may be expressed as

$$\pi^{(k+1)} = \pi^{(k)} T_{BGS} \qquad k = 0, 1, \cdots,$$

where

$$T_{BGS} = L(D - U)^{-1} \tag{3}$$

and $\pi^{(k)}$ is the approximate solution vector at the kth iteration. It is known that the spectral radius of T_{BGS} is equal to one; furthermore, π is the left eigenvector corresponding to the unit eigenvalue of T_{BGS}. The method of BGS will converge to the stationary vector for all $\pi^{(0)} \notin R(I - T_{BGS})$ (i.e, the initial approximation is not in the range of $(I - T_{BGS})$) if T_{BGS} does not have eigenvalues other than the unit eigenvalue on the unit circle (that is, if T_{BGS} is primitive). The asymptotic convergence rate of the BGS method depends on the magnitude of the subdominant eigenvalue of the iteration matrix, given by $\gamma(T_{BGS}) := \max\{|\lambda| \mid \lambda \in \sigma(T_{BGS}), \lambda \neq 1\}$. Here $\sigma(T_{BGS})$ denotes the set of eigenvalues of T_{BGS}. The smaller $\gamma(T_{BGS})$ is, the higher the asymptotic convergence rate of BGS.

Next, consider the coupling matrix $C^{(k)}$ computed in the aggregation step at the kth iteration of IAD, and whose ijth element is given by (see [11, p. 308])

$$c_{ij}^{(k)} = (\pi_i^{(k-1)}/\|\pi_i^{(k-1)}\|)P_{ij}e. \tag{4}$$

Here e is a column vector of ones and π_i denotes the ith subvector of π partitioned conformally with P in equation (1). For an NCD partitioning, the coupling matrix will be a perturbation of the identity matrix. The smallest eigenvalue of the (exact) coupling matrix indicates the inherent degree of ill-conditioning in the system. The disaggregation step of IAD uncouples the approximate solution vector using a BGS iteration to obtain an improved solution. We should also remark that IAD converges in a smaller number of iterations with a smaller degree of coupling [11, p. 340].

In Table 1, we present NCD partitionings for the Courtois matrix. While determining these partitionings we incremented the decomposability parameter ϵ by 0.025 times various powers of 10 as long as we had at least two subsets in the partitioning. Notice that for a given value of ϵ, the largest number of subsets in a computed partitioning is unique. There are 6 such partitionings and they are all nested [9] within each other for increasing values of ϵ. In Table 1, we report the smallest of such ϵ's. For the computed partitionings, the degree of coupling values increase from 0.0001 up to 0.4500. In the same table, we also provide the magnitude of the subdominant eigenvalue of the BGS iteration matrix given in equation (3) and the minimum eigenvalue of the exact coupling matrix in equation (4) using MATLAB. The exact coupling matrix C is computed using the exact stationary distribution π obtained with the Grassman-Taksar-Heyman (GTH) method [7].

Table 2 presents the last three NCD partitionings in Table 1 with their singletons grouped together. Observe that grouping singletons together sometimes has the effect of reducing the degree of coupling for a given NCD partitioning.

Results in Tables 1 and 2 for the Courtois matrix show that the degree of coupling and the smallest eigenvalue of the coupling matrix are inversely proportional. There is also correlation between the degree of coupling and the magnitude of the subdominant eigenvalue of the BGS iteration matrix. There seems to be a threshold for $\|E\|_\infty$ (0.15 for the Courtois matrix) below which $\gamma(T_{BGS})$ remains reasonably small. However, in between two NCD partitionings, the one that has the smaller $\|E\|_\infty$ may not necessarily have the smaller $\gamma(T_{BGS})$. For example, compare line 3 of Table 1 with line 2 of Table 2. Hence, the relationship is not monotonic. Finally, there is correlation between the magnitude of the subdominant eigenvalue of the BGS iteration matrix and the smallest eigenvalue of the coupling matrix. There seems to be a threshold this time for $\min(\sigma(C))$ (0.75 for the Courtois matrix) above which $\gamma(T_{BGS})$ remains reasonably small. However, an NCD partitioning that has the larger $\min(\sigma(C))$ may not necessarily have the smaller $\gamma(T_{BGS})$. Again, the relationship is not monotonic.

TABLE 1
NCD partitionings for the Courtois matrix.

Partition	ϵ	$\|E\|_\infty$	$\gamma(T_{BGS})$	$\min(\sigma(C))$
{1,2,3,4,5},{6,7,8}	$0.125e-3$	0.0001	0.0000	0.9998
{1,2,3},{4,5},{6,7,8}	$0.925e-3$	0.0010	0.0429	0.9985
{1},{2,3},{4,5},{6,7,8}	$0.150e+0$	0.1500	0.0539	0.7501
{1},{2,3},{4,5},{6},{7,8}	$0.250e+0$	0.4000	0.9956	0.4732
{1},{2,3},{4,5},{6},{7},{8}	$0.275e+0$	0.4500	0.9959	0.4000
{1},{2,3},{4},{5},{6},{7},{8}	$0.400e+0$	0.4500	0.9985	0.3007

TABLE 2
Effects of grouping singletons in NCD partitionings for the Courtois matrix.

Partition	ϵ	$\|E\|_\infty$	$\gamma(T_{BGS})$	$\min(\sigma(C))$
{2,3},{4,5},{7,8},{1,6}	$0.250e+0$	0.4000	0.9956	0.5745
{2,3},{4,5},{1,6,7,8}	$0.275e+0$	0.1499	0.0573	0.8762
{2,3},{1,4,5,6,7,8}	$0.400e+0$	0.1490	0.0002	0.8834

5. Conclusion. In this paper, we highlight a simple and useful algorithm that is able to compute NCD partitionings of Markov chains. The time and space complexity of the partitioning algorithm is linear in the number of nonzeros of the underlying Markov chain. In a recent paper [5], numerical experiments with BSOR and IAD on a test suite of large, sparse Markov chains have shown that there is merit in using the CC search algorithm especially when the chain is highly ill-conditioned. In practice, however, one may opt for a more balanced partitioning of blocks, and therefore end up using a partitioning with a larger value of the degree of coupling. Such a choice may depend on the order of the coupling matrix if IAD is the solver of choice, on the order of the largest diagonal block, and/or on available memory. In the same paper, it is shown that the partitioning time of TPABLO is substantial when compared with other partitioning techniques, and the partitionings TPABLO computes for a large number of parameter combinations provide no winning solver when used with BSOR or IAD on a representative subset of matrices in the test suite. Finally, grouping singletons in a given NCD partitioning into a separate subset may help because it is likely to reduce the value of the degree of coupling, and thereby improve the convergence rate of two-level iterative solvers.

References

[1] S. BAASE, *Computer Algorithms*, Addison-Wesley, Reading, MA, 1988.
[2] H. CHOI AND D. B. SZYLD, *Application of threshold partitioning of sparse matrices to Markov chains*, in the Proceedings of IEEE Int. Computer Performance and Dependability Symposium IPDS'96, Urbana, IL, 1996, pp. 158–165.
[3] P.-J. COURTOIS, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.
[4] T. DAYAR AND W. J. STEWART, *On the effects of using the Grassman-Taksar-Heyman method in iterative aggregation-disaggregation*, SIAM. J. Sci. Comput., 17 (1996), pp. 287–303.
[5] T. DAYAR AND W. J. STEWART, *Comparsion of partitioning techniques for two-level iterative solvers on large, sparse Markov chains*, SIAM. J. Sci. Comput., submitted for publication.
[6] I. S. DUFF, A. M. ERISMAN AND J. K. REID, *Direct Methods for Sparse Matrices*, Clarendon Press, Oxford, UK, 1986.
[7] W. K. GRASSMANN, M. I. TAKSAR AND D. P. HEYMAN, *Regenerative analysis and steady state distributions for Markov chains*, Oper. Res., 33 (1985), pp. 1107–1116.
[8] C. D. MEYER, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev., 31 (1989), pp. 240–272.
[9] M. E. SEZER AND D. D. ŠILJAK, *Nested ϵ-decompositions and clustering of complex systems*, Automatica, 22 (1986), pp. 321–331.
[10] M. E. SEZER AND D. D. ŠILJAK, *Nested epsilon decompositions of linear systems: weakly coupled and overlapping blocks*, SIAM J. Matrix. Anal. Appl., 12 (1991), pp. 521–533.
[11] W. J. STEWART, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, Princeton, NJ, 1994.
[12] W. J. STEWART AND W. WU, *Numerical experiments with iteration and aggregation for Markov chains*, ORSA J. Comput., 4 (1992), pp. 336–350.