UPDATING MARKOV CHAINS

AMY N. LANGVILLE* AND CARL D. MEYER†

1. Introduction. Suppose that the stationary distribution vector

$$\boldsymbol{\phi}^T = (\phi_1, \phi_2, \ldots, \phi_m)$$

for an m-state homogeneous irreducible Markov chain with transition probability matrix $\mathbf{Q}_{m\times m}$ is known, but the chain requires updating by altering some of its transition probabilities or by adding or deleting some states. Suppose that the updated transition probability matrix $\mathbf{P}_{n\times n}$ is also irreducible. The updating problem is to compute the updated stationary distribution $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \ldots, \pi_n)$ for $\mathbf{P}$ by somehow using the components in $\boldsymbol{\phi}^T$ to produce $\boldsymbol{\pi}^T$ with less effort than that required by starting from scratch.

2. The Power Method. For the simple case in which the updating process calls for perturbing transition probabilities in $\mathbf{Q}$ to produce the updated matrix $\mathbf{P}$ without creating or destroying states, restarting the power method is a possible updating technique. In other words, simply iterate with the new transition matrix but use the old stationary distribution as the initial vector

$$\mathbf{x}_{j+1}^T = \mathbf{x}_j^T\mathbf{P} \quad \text{with} \quad \mathbf{x}_0^T = \boldsymbol{\phi}^T. \tag{1}$$

Will this produce an acceptably accurate approximation to $\boldsymbol{\pi}^T$ in fewer iterations than are required when an arbitrary initial vector is used? To some degree this is true, but intuition generally overestimates the extent, even when updating produces a $\mathbf{P}$ that is close to $\mathbf{Q}$. For example, if the entries of $\mathbf{P}-\mathbf{Q}$ are small enough to ensure that each component π_i agrees with ϕ_i in the first significant digit, and if the goal is to compute the update $\boldsymbol{\pi}^T$ to twelve significant places by using (1), then about $11/R$ iterations are required, whereas starting from scratch with an initial vector containing no significant digits of accuracy requires about $12/R$ iterations, where $R = -\log_{10}|\lambda_2|$ is the asymptotic rate of convergence with λ_2 being the subdominant eigenvalue of $\mathbf{P}$. In other words, the effort is reduced by about 8% for each correct significant digit that is built into $\mathbf{x}_0^T$ [22]. In general, the restarted power method is not particularly effective as an updating technique, even when the updates represent small perturbations.

3. Rank-One Updating. When no states are added or deleted, the updating problem can be formulated in terms of updating $\mathbf{Q}$ one row at a time by adapting the Sherman–Morrison rank-one updating formula [29] to the singular matrix $\mathbf{A} = \mathbf{I}-\mathbf{Q}$. The mechanism for doing this is by means of the group inverse $\mathbf{A}^\#$ for $\mathbf{A}$, which is often involved in questions concerning Markov chains [6, 9, 11, 14, 23, 25, 27, 28, 31, 32, 38]. $\mathbf{A}^\#$ is the unique matrix satisfying $\mathbf{A}\mathbf{A}^\#\mathbf{A} = \mathbf{A}$, $\mathbf{A}^\#\mathbf{A}\mathbf{A}^\# = \mathbf{A}^\#$, and $\mathbf{A}\mathbf{A}^\# = \mathbf{A}^\#\mathbf{A}$.

THEOREM 3.1. [31]. *If the ith row $\mathbf{q}^T$ of $\mathbf{Q}$ is updated to produce $\mathbf{p}^T = \mathbf{q}^T - \boldsymbol{\delta}^T$, the ith row of $\mathbf{P}$, and if $\boldsymbol{\phi}^T$ and $\boldsymbol{\pi}^T$ are the respective stationary probability distribu-*

*Department of Mathematics, College of Charleston, Charleston, SC 29424, (langvillea@cofc.edu).
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, (meyer@ncsu.edu).

tions of $\mathbf{Q}$ *and* $\mathbf{P}$*, then then* $\boldsymbol{\pi}^T = \boldsymbol{\phi}^T - \boldsymbol{\epsilon}^T$*, where*

$$\boldsymbol{\epsilon}^T = \left[\frac{\phi_i}{1 + \boldsymbol{\delta}^T \mathbf{A}^{\#}_{*i}}\right] \boldsymbol{\delta}^T \mathbf{A}^{\#} \qquad (\mathbf{A}^{\#}_{*i} = \textit{the } i\textit{th column of } \mathbf{A}^{\#}). \tag{2}$$

Multiple row updates to $\mathbf{Q}$ *are accomplished by sequentially applying this formula one row at a time, which means that the group inverse must be sequentially updated. The formula for updating* $(\mathbf{I} - \mathbf{Q})^{\#}$ *to* $(\mathbf{I} - \mathbf{P})^{\#}$ *is as follows:*

$$(\mathbf{I} - \mathbf{P})^{\#} = \mathbf{A}^{\#} + \mathbf{e}\boldsymbol{\epsilon}^T \left[\mathbf{A}^{\#} - \gamma \mathbf{I}\right] - \frac{\mathbf{A}^{\#}_{*i}\boldsymbol{\epsilon}^T}{\phi_i}, \quad \textit{where} \quad \gamma = \frac{\boldsymbol{\epsilon}^T \mathbf{A}^{\#}_{*i}}{\phi_i}. \tag{3}$$

If more than just one or two rows are involved, then Theorem 3.1 is not computationally efficient. If every row needs to be touched, then using (2) together with (3) requires $O(n^3)$ floating point operations, which is comparable to the cost of starting from scratch. Other updating formulas exist [9, 12, 16, 19, 36], but all are variations of a Sherman–Morrison type of formula, and all are $O(n^3)$ algorithms for a general update. Moreover, rank-one updating techniques are not easily adapted to handle the creation or destruction of states.

4. Aggregation. Consider an irreducible n-state Markov chain whose state space has been partitioned into k disjoint groups $\mathcal{S} = G_1 \cup G_2 \cup \cdots \cup G_k$ with associated transition probability matrix

$$\mathbf{P}_{n\times n} = \begin{array}{c} \\ G_1 \\ G_2 \\ \vdots \\ G_k \end{array} \begin{array}{c} \begin{array}{cccc} G_1 & G_2 & \cdots & G_k \end{array} \\ \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1k} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{k1} & \mathbf{P}_{k2} & \cdots & \mathbf{P}_{kk} \end{pmatrix} \end{array} \quad \text{(square diagonal blocks)}. \tag{4}$$

This *parent* chain induces k smaller Markov chains called *censored chains*. The censored chain associated with G_i is defined to be the Markov process that records the location of the parent chain only when the parent chain visits states in G_i. Visits to states outside of G_i are ignored. The transition probability matrix for the ith censored chain is the ith *stochastic complement* [26] defined by

$$\mathbf{S}_i = \mathbf{P}_{ii} + \mathbf{P}_{i\star}(\mathbf{I} - \mathbf{P}^{\star}_i)^{-1}\mathbf{P}_{\star i}, \tag{5}$$

in which $\mathbf{P}_{i\star}$ and $\mathbf{P}_{\star i}$ are, respectively, the ith row and the ith column of blocks with $\mathbf{P}_{ii}$ removed, and $\mathbf{P}^{\star}_i$ is the principal submatrix of $\mathbf{P}$ obtained by deleting the ith row and ith column of blocks. For example, if $\mathcal{S} = G_1 \cup G_2$, then the respective transition matrices for the two censored chains are the two stochastic complements

$$\mathbf{S}_1 = \mathbf{P}_{11} + \mathbf{P}_{12}(\mathbf{I} - \mathbf{P}_{22})^{-1}\mathbf{P}_{21} \quad \text{and} \quad \mathbf{S}_2 = \mathbf{P}_{22} + \mathbf{P}_{21}(\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{P}_{12}.$$

In general, if the stationary distribution for $\mathbf{P}$ is $\boldsymbol{\pi}^T = (\,\boldsymbol{\pi}_1^T \,|\, \boldsymbol{\pi}_2^T \,|\, \cdots \,|\, \boldsymbol{\pi}_k^T\,)$ (partitioned conformably with $\mathbf{P}$), then the *ith censored distribution* (the stationary distribution for $\mathbf{S}_i$) is

$$\mathbf{s}_i^T = \frac{\boldsymbol{\pi}_i^T}{\boldsymbol{\pi}_i^T \mathbf{e}}, \qquad \text{where } \mathbf{e} \text{ is an appropriately sized column of ones [26].} \tag{6}$$

For aperiodic chains, the jth component of $\mathbf{s}_i^T$ is the limiting conditional probability of being in the jth state of group G_i given that the process is somewhere in G_i.

Each group G_i is compressed into a single state in a smaller k-state aggregated chain by squeezing the original transition matrix $\mathbf{P}$ down to an *aggregated transition matrix*

$$\mathbf{A}_{k\times k} = \begin{pmatrix} \mathbf{s}_1^T\mathbf{P}_{11}\mathbf{e} & \cdots & \mathbf{s}_1^T\mathbf{P}_{1k}\mathbf{e} \\ \vdots & \ddots & \vdots \\ \mathbf{s}_k^T\mathbf{P}_{k1}\mathbf{e} & \cdots & \mathbf{s}_k^T\mathbf{P}_{kk}\mathbf{e} \end{pmatrix}, \tag{7}$$

which is stochastic and irreducible whenever $\mathbf{P}$ is [26]. For aperiodic chains, transitions between states in the aggregated chain defined by $\mathbf{A}$ correspond to transitions between groups G_i in the parent chain when the parent chain is in equilibrium. The remarkable feature of aggregation is that it allows the parent chain to be decomposed into k small censored chains that can be independently solved, and the resulting censored distributions $\mathbf{s}_i^T$ can be combined with the stationary distribution of $\mathbf{A}$ to construct the parent stationary distribution $\boldsymbol{\pi}^T$. This is the aggregation theorem.

THEOREM 4.1. (*the aggregation theorem* [26]) *If* $\mathbf{P}$ *is the block-partitioned transition probability matrix* (4) *for an irreducible* n*-state Markov chain whose stationary probability distribution is*

$$\boldsymbol{\pi}^T = (\,\boldsymbol{\pi}_1^T \,|\, \boldsymbol{\pi}_2^T \,|\, \cdots \,|\, \boldsymbol{\pi}_k^T\,) \quad \textit{(partitioned conformably with } \mathbf{P}\textit{)},$$

and if $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ *is the stationary distribution for the aggregated chain defined by the matrix* $\mathbf{A}_{k\times k}$ *in* (7)*, then* $\alpha_i = \boldsymbol{\pi}_i^T\mathbf{e}$*, and the stationary distribution for* $\mathbf{P}$ *is*

$$\boldsymbol{\pi}^T = \left(\alpha_1\mathbf{s}_1^T \,|\, \alpha_2\mathbf{s}_2^T \,|\, \cdots \,|\, \alpha_k\mathbf{s}_k^T\right),$$

where $\mathbf{s}_i^T$ *is the censored distribution for the stochastic complement* $\mathbf{S}_i$ *in* (5).

5. Approximate Updating by Aggregation. Aggregation as presented in Theorem 4.1 is mathematically elegant but numerically inefficient because costly inversions are embedded in the stochastic complements (5) that are required to produce the censored distributions $\mathbf{s}_i^T$. Consequently, the approach is to derive computationally cheap estimates of the censored distributions as described below.

In many large-scale problems the effects of updating are localized. That is, not all stationary probabilities are equally affected—some changes may be significant while others are hardly perceptible—e.g., this is generally true in applications such as Google's PageRank Problem [21] in which the stationary probabilities obey a power-law distribution (discussed in section 7.2).

Partition the state space of the updated chain as $\mathcal{S} = G \cup \overline{G}$, where G contains the states that are most affected by updating along with any new states created by updating—techniques for determining these states are discussed in section 7. Some nearest neighbors of newly created states might also go into G. Partition the updated (and reordered) transition matrix $\mathbf{P}$ as

$$\mathbf{P} = \left(\begin{array}{c|c|c|c} p_{11} & \cdots & p_{1g} & \mathbf{P}_{1\star} \\ \hline \vdots & \ddots & \vdots & \vdots \\ \hline p_{g1} & \cdots & p_{gg} & \mathbf{P}_{g\star} \\ \hline \mathbf{P}_{\star 1} & \cdots & \mathbf{P}_{\star g} & \mathbf{P}_{22} \end{array}\right) = \begin{array}{c} \\ G \\ \overline{G} \end{array} \begin{array}{c} \begin{array}{cc} G & \overline{G} \end{array} \\ \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \end{array}, \tag{8}$$

where

$$\mathbf{P}_{11} = \begin{pmatrix} p_{11} & \cdots & p_{1g} \\ \vdots & \ddots & \vdots \\ p_{g1} & \cdots & p_{gg} \end{pmatrix}, \quad \mathbf{P}_{12} = \begin{pmatrix} \mathbf{P}_{1\star} \\ \vdots \\ \mathbf{P}_{g\star} \end{pmatrix}, \quad \text{and} \quad \mathbf{P}_{21} = (\mathbf{P}_{\star 1} \cdots \mathbf{P}_{\star g}).$$

Let $\boldsymbol{\phi}^T$ and $\boldsymbol{\pi}^T$ be the respective stationary distributions of the pre- and post-updated chains, and let if $\overline{\boldsymbol{\phi}}^T$ and $\overline{\boldsymbol{\pi}}^T$ contain the respective stationary probabilities from $\boldsymbol{\phi}^T$ and $\boldsymbol{\pi}^T$ that correspond to the states in $\overline{G}$. The stipulation that $\overline{G}$ contains the nearly unaffected states translates to saying that

$$\overline{\boldsymbol{\pi}}^T \approx \overline{\boldsymbol{\phi}}^T.$$

When viewed as a partitioned matrix with $g+1$ diagonal blocks, the first g diagonal blocks in $\mathbf{P}$ are 1×1, and the lower right-hand block is the $(n-g) \times (n-g)$ matrix $\mathbf{P}_{22}$ that is associated with the states in $\overline{G}$. The stochastic complements in $\mathbf{P}$ are

$$\mathbf{S}_1 = \cdots = \mathbf{S}_g = 1, \quad \text{and} \quad \mathbf{S}_{g+1} = \mathbf{P}_{22} + \mathbf{P}_{21}\big(\mathbf{I} - \mathbf{P}_{11}\big)^{-1}\mathbf{P}_{12}.$$

Consequently, the aggregated transition matrix (7) becomes

$$\mathbf{A} = \left(\begin{array}{c|c|c|c} p_{11} & \cdots & p_{1g} & \mathbf{P}_{1\star}\mathbf{e} \\ \hline \vdots & \ddots & \vdots & \vdots \\ \hline p_{g1} & \cdots & p_{gg} & \mathbf{P}_{g\star}\mathbf{e} \\ \hline \mathbf{s}^T\mathbf{P}_{\star 1} & \cdots & \mathbf{s}^T\mathbf{P}_{\star g} & \mathbf{s}^T\mathbf{P}_{22}\mathbf{e} \end{array}\right)_{(g+1)\times(g+1)} \tag{9}$$

$$= \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12}\mathbf{e} \\ \mathbf{s}^T\mathbf{P}_{21} & \mathbf{s}^T\mathbf{P}_{22}\mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12}\mathbf{e} \\ \mathbf{s}^T\mathbf{P}_{21} & 1 - \mathbf{s}^T\mathbf{P}_{21}\mathbf{e} \end{pmatrix},$$

where $\mathbf{s}^T$ is the censored distribution derived from the only significant stochastic complement $\mathbf{S} = \mathbf{S}_{g+1}$. If the stationary distribution for $\mathbf{A}$ is

$$\boldsymbol{\alpha}^T = \big(\alpha_1, \ldots, \alpha_g, \alpha_{g+1}\big),$$

then Theorem 4.1 says that the stationary distribution for $\mathbf{P}$ is

$$\boldsymbol{\pi}^T = (\pi_1, \ldots \pi_g \,|\, \pi_{g+1}, \ldots, \pi_n) = \big(\pi_1, \ldots, \pi_g \,|\, \overline{\boldsymbol{\pi}}^T\big) = \big(\alpha_1, \ldots, \alpha_g \,|\, \overline{\boldsymbol{\pi}}^T\big) \tag{10}$$

Since $\mathbf{s}_i^T = \boldsymbol{\pi}_i^T/\boldsymbol{\pi}_i^T\mathbf{e}$, and since $\overline{\boldsymbol{\pi}}^T \approx \overline{\boldsymbol{\phi}}^T$, it follows that

$$\tilde{\mathbf{s}}^T = \frac{\overline{\boldsymbol{\phi}}^T}{\overline{\boldsymbol{\phi}}^T\mathbf{e}} \approx \mathbf{s}^T \tag{11}$$

is a good approximation to $\mathbf{s}^T$ that is available from the pre-updated distribution. Using this in (9) produces an approximate aggregated transition matrix

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12}\mathbf{e} \\ \tilde{\mathbf{s}}^T\mathbf{P}_{21} & 1 - \tilde{\mathbf{s}}^T\mathbf{P}_{21}\mathbf{e} \end{pmatrix}. \tag{12}$$

Notice that

$$\mathbf{A} - \widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{\delta}^T\mathbf{P}_{21} & -\boldsymbol{\delta}^T\mathbf{P}_{21}\mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\delta}^T \end{pmatrix} \mathbf{P}_{21}(\mathbf{I} \mid -\mathbf{e}), \quad \text{where} \quad \boldsymbol{\delta}^T = \mathbf{s}^T - \tilde{\mathbf{s}}^T.$$

Consequently, $\mathbf{A} - \widetilde{\mathbf{A}}$ and $\mathbf{s}^T - \tilde{\mathbf{s}}^T$ are of the same order of magnitude, so the stationary distribution $\widetilde{\boldsymbol{\alpha}}^T$ of $\widetilde{\mathbf{A}}$ can provide a good approximation to $\boldsymbol{\alpha}^T$, the stationary distribution of $\mathbf{A}$. That is,

$$\widetilde{\boldsymbol{\alpha}}^T = (\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_g, \tilde{\alpha}_{g+1}) \approx \big(\alpha_1, \ldots, \alpha_g, \alpha_{g+1}\big) = \boldsymbol{\alpha}^T.$$

Use $\tilde{\alpha}_i \approx \alpha_i$ for $1 \leq i \leq g$ in (10) along with $\overline{\boldsymbol{\pi}}^T \approx \overline{\boldsymbol{\phi}}^T$ to obtain the approximation

$$\boldsymbol{\pi}^T \approx \widetilde{\boldsymbol{\pi}}^T = \left(\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_g \,|\, \overline{\boldsymbol{\phi}}^T\right). \tag{13}$$

Thus an approximate updated distribution is obtained. The degree to which this approximation is accurate clearly depends on the degree to which $\overline{\boldsymbol{\pi}}^T \approx \overline{\boldsymbol{\phi}}^T$. If (13) does not provide the desired accuracy, it can be viewed as the first step in an iterative aggregation scheme described below that performs remarkably well.

6. Updating by Iterative Aggregation. Iterative aggregation as described in [38] is not a general-purpose technique, because it usually does not work for chains that are not nearly uncoupled. However, iterative aggregation can be adapted to the updating problem, and these variations work extremely well, even for chains that are not nearly uncoupled. This is in part due to the fact that the approximate aggregation matrix (12) differs from the exact aggregation matrix (9) in only one row. Our iterative aggregation updating algorithm is described below.

Assume that the stationary distribution $\boldsymbol{\phi}^T = (\phi_1, \phi_2, \ldots, \phi_m)$ for some irreducible Markov chain $\mathcal{C}$ is already known, perhaps from prior computations, and suppose that $\mathcal{C}$ needs to be updated. As in earlier sections, let the transition probability matrix and stationary distribution for the updated chain be denoted by $\mathbf{P}$ and $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \ldots, \pi_n)$, respectively. The updated matrix $\mathbf{P}$ is assumed to be irreducible. It is important to note that m is not necessarily equal to n because the updating process allows for the creation or destruction of states as well as the alteration of transition probabilities.

The Iterative Aggregation Updating Algorithm

Initialization

i. Partition the states of the updated chain as $\mathcal{S} = G \cup \overline{G}$ and reorder $\mathbf{P}$ as described in (8)

ii. $\overline{\boldsymbol{\phi}}^T \longleftarrow$ the components from $\boldsymbol{\phi}^T$ that correspond to the states in $\overline{G}$

iii. $\mathbf{s}^T \longleftarrow \overline{\boldsymbol{\phi}}^T/(\overline{\boldsymbol{\phi}}^T\mathbf{e})$ (an initial approximate censored distribution)

Iterate until convergence

1. $\mathbf{A} \longleftarrow \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12}\mathbf{e} \\ \mathbf{s}^T\mathbf{P}_{21} & 1 - \mathbf{s}^T\mathbf{P}_{21}\mathbf{e} \end{pmatrix}_{(g+1)\times(g+1)}$ $(\, g = |G| \,)$
2. $\boldsymbol{\alpha}^T \longleftarrow (\alpha_1, \alpha_2, \ldots, \alpha_g, \alpha_{g+1})$ (the stationary distribution of $\mathbf{A}$)
3. $\boldsymbol{\chi}^T \longleftarrow \left(\alpha_1, \alpha_2, \ldots, \alpha_g \,|\, \alpha_{g+1}\mathbf{s}^T\right)$
4. $\boldsymbol{\psi}^T \longleftarrow \boldsymbol{\chi}^T\mathbf{P}$ (see note following the algorithm)
5. If $\|\boldsymbol{\psi}^T - \boldsymbol{\chi}^T\| < \tau$ for a given tolerance τ, then quit—else $\mathbf{s}^T \longleftarrow \boldsymbol{\psi}^T/\boldsymbol{\psi}^T\mathbf{e}$ and go to step 1

Note concerning step 4. Step 4 is necessary because the vector $\boldsymbol{\chi}^T$ generated in step 3 is a fixed point in the sense that if Step 4 is omitted and the process is restarted using $\boldsymbol{\chi}^T$ instead of $\boldsymbol{\psi}^T$, then the same $\boldsymbol{\chi}^T$ is simply reproduced at Step 3 on each subsequent iteration. Step 4 has two purposes—it moves the iterate off the fixed point while simultaneously contributing to the convergence process. That is, the $\boldsymbol{\psi}^T$ resulting from step 4 can be used to restart the algorithm as well as produce a better approximation because applying a power step makes small progress toward the stationary solution. In the past, some authors [38] have used Gauss–Seidel in place of the power method at Step 4.

While precise rates of convergence for general iterative aggregation algorithms are difficult to articulate, the specialized nature of our iterative aggregation updating algorithm allows us to easily establish its rate of convergence. The following theorem shows that this rate is directly dependent on how fast the powers of the one significant stochastic complement $\mathbf{S} = \mathbf{P}_{22} + \mathbf{P}_{21}(\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{P}_{12}$ converge. In other words, since $\mathbf{S}$ is an irreducible stochastic matrix, the rate of convergence is completely dictated by the magnitude and Jordan structure of the largest subdominant eigenvalue of $\mathbf{S}$.

THEOREM 6.1. [22] *The iterative aggregation updating algorithm defined above converges to the stationary distribution* $\boldsymbol{\pi}^T$ *of* $\mathbf{P}$ *for all partitions* $\mathcal{S} = G \cup \overline{G}$. *The rate at which the iterates converge to* $\boldsymbol{\pi}^T$ *is exactly the rate at which the powers* $\mathbf{S}^n$ *converge, which is governed by the magnitude and Jordan structure of largest subdominant eigenvalue* $\lambda_2(\mathbf{S})$ *of* $\mathbf{S}$. *If* $\lambda_2(\mathbf{S})$ *is real and simple, then the asymptotic rate of convergence is* $R = -\log_{10}|\lambda_2(\mathbf{S})|$.

7. Determining The Partition. The iterative aggregation updating algorithm is globally convergent, and it never requires more iterations than the power method to attain a given level of convergence [17]. However, iterative aggregation clearly requires more work per iteration than the power method. One iteration of iterative aggregation requires forming the aggregation matrix, solving for its stationary vector, and executing one power iteration. The key to realizing an improvement in iterative aggregation over the power method rests in properly choosing the partition $\mathcal{S} = G \cup \overline{G}$. As Theorem 6.1 shows, good partitions are precisely those that yield a stochastic complement $\mathbf{S} = \mathbf{P}_{22} + \mathbf{P}_{21}(\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{P}_{12}$ whose subdominant eigenvalue $\lambda_2(\mathbf{S})$ is small in magnitude.

Experience indicates that as $|G| = g$ (the size of $\mathbf{P}_{11}$) becomes larger, iterative aggregation tends to converge in fewer iterations. But as g becomes larger, each iteration requires more work, so the trick is to strike an acceptable balance. A small g that significantly reduces $|\lambda_2(\mathbf{S})|$ is the ideal situation.

Even for moderately sized problems there is an extremely large number of possible partitions, but there are some useful heuristics that can help guide the choice of G that will produce reasonably good results. For example, a relatively simple approach is to take G to be the set of all states "near" the updates, where "near" might be measured in a graph theoretic sense or else by transient flow [7] (i.e., using the magnitude of entries of $\mathbf{x}_{j+1}^T = \mathbf{x}_j^T\mathbf{P}$ after j iterations, where j is small, say 5 or 10). In the absence of any other information, this naive strategy is at least a good place to start. However, there are usually additional options that lead to even better "G-sets," and some of these are described below.

7.1. Partitioning by differing time scales. In most aperiodic chains, evolution is not at a uniform rate, and consequently most iterative techniques, including the power method, often spend the majority of the time in resolving a small number of components—the slow evolving states. The slow states can be isolated either by monitoring the process for a few iterations or by theoretical means [22]. If the slow states are placed in G while the faster-converging states are lumped into $\overline{G}$, then the iterative aggregation algorithm concentrates its effort on resolving the smaller number of slow-converging states.

In loose terms, the effect of steps 1–3 in the iterative aggregation algorithm is essentially to make progress toward achieving a steady state for a smaller chain consisting of just the slow states in G together with one additional lumped state that accounts for all fast states in $\overline{G}$. The power iteration in step 4 moves the entire process ahead on a global basis, so if the slow states in G are substantially resolved by the relatively cheaper steps 1—3, then not many of the more costly global power steps are required to push the entire chain toward its global equilibrium. This is the essence of the original Simon–Ando idea first proposed in 1961 and explained and analyzed in [26, 37]. As $g = |G|$ becomes smaller relative to n, steps 1–3 become significantly cheaper to execute, and the process converges rapidly in both iteration count and wall-clock time. Examples and reports on experiments are given in [22].

In some applications the slow states are particularly easy to identify because they are the ones having the larger stationary probabilities. This is a particularly nice state of affairs for the updating problem because we have the stationary probabilities from the prior period at our disposal, and thus all we have to do to construct a good G-set is to include the states with prior large stationary probabilities and throw in the states that were added or updated along with a few of their nearest neighbors. Clearly, this is an advantage only when there are just a few "large" states. However, it turns out that this is a characteristic feature of scale-free networks with power-law distributions [1, 2, 5, 10] discussed below.

7.2. Scale-free networks. A scale-free networks with a power-law distribution is a network in which the number of nodes $n(l)$ having l edges (possibly directed) is proportional to l^{-k} where k is a constant that does not change as the network expands (hence the term "scale-free"). In other words, the distribution of nodal degrees obeys a *power-law distribution* in the sense that

$$P[\deg(N) = d] \propto \frac{1}{d^k} \qquad \text{for some } k > 1 \qquad (\propto \text{ means "proportional to"}).$$

A Markov chain with a power-law distribution is a chain in which there are relatively very few states that have a significant stationary probability while the overwhelming majority of states have nearly negligible stationary probabilities. Google's

PageRank application [21, 22] is an important example. Consequently, when the stationary probabilities are plotted in order of decreasing magnitude, the resulting graph has a pronounced "L-shape" with an extremely sharp bend. It is this characteristic "L-shape" that reveals a near optimal partition $\mathcal{S} = G \cup \overline{G}$ for the iterative aggregation updating algorithm presented in section 6. Experiments indicate that the size of the G-set used in our iterative aggregation updating algorithm is nearly optimal around a point that is just to the right-hand side of the pronounced bend in the L-curve. In other words, an apparent method for constructing a reasonably good partition $\mathcal{S} = G \cup \overline{G}$ for the iterative aggregation updating algorithm is as follows.

1. First put all new states and states with newly created or destroyed connections (perhaps along with some of their nearest neighbors) into G.
2. Add other states that remain after the update in order of the magnitude of their prior stationary probabilities up to the point where these stationary probabilities level off.

Of course, there is some subjectiveness to this strategy. However, the leveling-off point is relatively easy to discern in distributions having a sharply defined bend in the L-curve, and only distributions that gradually die away or do not conform to a power law are problematic. If, when ordered by magnitude, the stationary probabilities

$$\pi(1) \geq \pi(2) \geq \cdots \geq \pi(n)$$

for an irreducible chain conform to a power-law distribution so that there are constants $\alpha > 0$ and $k > 0$ such that $\pi(i) \approx \alpha i^{-k}$, then the "leveling-off point" i_{level} can be taken to be the smallest value for which $|d\pi(i)/di| \approx \epsilon$ for some user-defined tolerance ϵ. That is, $i_{level} \approx (k\alpha/\epsilon)^{1/k+1}$. This provides a rough estimate of g_{opt}, the optimal size of G, but empirical evidence suggests that better estimates require a scaling factor $\sigma(n)$ that accounts for the size of the chain; i.e.,

$$g_{opt} \approx \sigma(n) \left(\frac{k\alpha}{\epsilon}\right)^{1/k+1} \approx \sigma(n) \left(\frac{k\pi(1)}{\epsilon}\right)^{1/k+1}.$$

More research and testing is needed to resolve some of these issues.

REFERENCES

[1] Albert-Laszlo Barabasi. *Linked: The New Science of Networks*. Plume, 2003.

[2] Albert-Laszlo Barabasi, Reka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, 281:69–77, 2000.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 33:107–117, 1998.

[4] Sergey Brin, Lawrence Page, R. Motwami, and Terry Winograd. The PageRank citation ranking: bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *The Ninth International WWW Conference*, May 2000. `http://www9.org/w9cdrom/160/160.html`.

[6] Steven Campbell and Carl D. Meyer. *Generalized Inverses of Linear Transformations*. Pitman, San Francisco, 1979.

[7] Steve Chien, Cynthia Dwork, Ravi Kumar, and D. Sivakumar. Towards exploiting link evolution. In *Workshop on algorithms and models for the Web graph*, 2001.

[8] Grace E. Cho and Carl D. Meyer. Markov chain sensitivity measured by mean first passage times. *Linear Algebra and its Applications*, 313:21–28, 2000.

[9] Grace E. Cho and Carl D. Meyer. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1–3):137–150, 2001.
[10] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *The European Physical Journal B*, 38:239–243, 2004.
[11] Robert E. Funderlic and Carl D. Meyer. Sensitivity of the stationary distribution vector for an ergodic Markov chain. *Linear Algebra and its Applications*, 76:1–17, 1986.
[12] Robert E. Funderlic and Robert J. Plemmons. Updating **LU** factorizations for computing stationary distributions. *SIAM Journal on Algebraic and Discrete Methods*, 7(1):30–42, 1986.
[13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.
[14] Gene H. Golub and Carl D. Meyer. Using the QR factorization and group inverse to compute, differentiate and estimate the sensitivity of stationary probabilities for Markov chains. *SIAM Journal on Algebraic and Discrete Methods*, 17:273–281, 1986.
[15] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.
[16] Jeffrey J. Hunter. Stationary distributions of perturbed Markov chains. *Linear Algebra and its Applications*, 82:201–214, 1986.
[17] Ilse C. F. Ipsen and Steve Kirkland. Convergence analysis of an improved PageRank algorithm. Technical Report CRSC-TR04-02, North Carolina State University, 2004.
[18] Ilse C. F. Ipsen and Carl D. Meyer. Uniform stability of Markov chains. *SIAM Journal on Matrix Analysis and Applications*, 15(4):1061–1074, 1994.
[19] John G. Kemeny and Laurie J. Snell. *Finite Markov Chains*. D. Van Nostrand, New York, 1960.
[20] Amy N. Langville and Carl D. Meyer. A survey of eigenvector methods of web information retrieval. *SIAM Rev.*, 47(1):135–161, 2005.
[21] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.
[22] Amy N. Langville and Carl D. Meyer. Updating the stationary vector of an irreducible Markov chain with an eye on Google's PageRank. *SIAM Journal on Matrix Analysis and Applications*, 27:968–987, 2006.
[23] Carl D. Meyer. The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev.*, 17:443–464, 1975.
[24] Carl D. Meyer. The condition of a finite Markov chain and perturbation bounds for the limiting probabilities. *SIAM Journal on Algebraic and Discrete Methods*, 1:273–283, 1980.
[25] Carl D. Meyer. Analysis of finite Markov chains by group inversion techniques. *Recent Applications of Generalized Inverses, Research Notes in Mathematics, Pitman, Ed. S. L. Campbell*, 66:50–81, 1982.
[26] Carl D. Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272, 1989.
[27] Carl D. Meyer. The character of a finite Markov chain. *Linear Algebra, Markov Chains, and Queueing Models, IMA Volumes in Mathematics and its Applications, Ed., Carl D. Meyer and Robert J. Plemmons, Springer-Verlag*, 48:47–58, 1993.
[28] Carl D. Meyer. Sensitivity of the stationary distribution of a Markov chain. *SIAM Journal on Matrix Analysis and Applications*, 15(3):715–728, 1994.
[29] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2000.
[30] Carl D. Meyer and Robert J. Plemmons. *Linear Algebra, Markov Chains, and Queueing Models*. Springer-Verlag, 1993.
[31] Carl D. Meyer and James M. Shoaf. Updating finite Markov chains by using techniques of group matrix inversion. *Journal of Statistical Computation and Simulation*, 11:163–181, 1980.
[32] Carl D. Meyer and G. W. Stewart. Derivatives and perturbations of eigenvectors. *SIAM J. Numer. Anal.*, 25:679–691, 1988.
[33] Cleve Moler. The world's largest matrix computation. *Matlab News and Notes*, pages 12–13, October 2002.
[34] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.
[35] Eugene Seneta. *Non-negative matrices and Markov chains*. Springer-Verlag, 1981.
[36] Eugene Seneta. Sensivity analysis, ergodicity coefficients, and rank-one updates for finite Markov chains. In William J. Stewart, editor, *Numerical Solutions of Markov Chains*, pages 121–129, 1991.

[37] Herbert A. Simon and Albert Ando. Aggregation of variables in dynamic systems. *Econometrica*, 29:111–138, 1961.
[38] William J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.
[39] Twelfth International World Wide Web Conference. *Extrapolation Methods for Accelerating PageRank Computations*, 2003.