# A COMBINATORIAL APPROACH TO NEARLY UNCOUPLED MARKOV CHAINS

A thesis submitted in partial fulfillment

of the requirements for the degree of

Philosophiæ Doctor

By

Ryan Tifenbach

Research Supervisor: Prof. Steve Kirkland

Head of Department: Prof. Douglas Leith

Hamilton Institute

National University of Ireland Maynooth

September 2011

# Abstract

A discrete-time Markov chain on a state space $\mathcal{S}$ is a sequence of random variables $X = \{x_0, x_1, \ldots\}$ that take on values in $\mathcal{S}$. A Markov chain is a model of a system which changes or evolves over time; the random variable $x_t$ is the state of the system at time $t$.

A subset $\mathcal{E} \subseteq \mathcal{S}$ is referred to as an almost invariant aggregate if whenever $x_t \in \mathcal{E}$, then with high probability $x_{t+1} \in \mathcal{E}$, as well. That is, if there is a small positive value $\epsilon$ such that if $x_t \in \mathcal{E}$ then the probability that $x_{t+1} \notin \mathcal{E}$ is less than or equal to $\epsilon$, then $\mathcal{E}$ is an almost invariant aggregate. If $\mathcal{E}$ is such an aggregate and $x_t \in \mathcal{E}$, then the probability that $x_{t+1}, \ldots, x_{t+s} \in \mathcal{E}$ is at least $(1-\epsilon)^s$. A Markov chain tends to remain within its almost invariant aggregates (if it possesses any) for long periods of time.

We refer to the Markov chain $X$ as nearly uncoupled (with respect to some positive $\epsilon$) if its associated state space contains two or more disjoint almost invariant aggregates. Nearly uncoupled Markov chains are characterised by long periods of

relatively constant behaviour, punctuated by occasional drastic changes in state.

We present a series of algorithms intended to construct almost invariant aggregates of a given Markov chain. These algorithms are iterative processes which utilise a concept known as the stochastic complement. The stochastic complement is a method by which a Markov chain on a state space $\mathcal{S}$ can be reduced to a random process on a proper subset $\mathcal{S}' \subseteq \mathcal{S}$, while preserving many of the algebraic properties of the original Markov chain.

We pay special attention to the reversible case. A Markov chain is reversible if it is symmetric in time – by which we mean that if we were to reverse the order of the variables $x_1, \ldots, x_t$, for some relatively large $t$, the resulting process would be essentially indistinguishable from the original Markov chain.

# Acknowledgements

I want to thank Steve Kirkland for his generous support during my graduate studies. His help far surpassed what might normally be expected of an advisor and his assistance at every step of the process has been invaluable.

As well, I would like to acknowledge the support of the faculty members of the Hamilton Institute at the National University of Ireland Maynooth and the Department of Mathematics and Statistics at the University of Regina. In particular, I want to thank Shaun Fallat, Doug Farenick and Oliver Mason for their kind assistance.

Finally, I would like to thank Jessica Hampton, my wife, for her loving and ceaseless support throughout my academic career.

This work is dedicated to my son, Ivan Hampton, who inspires me more than I can express.

# A note concerning the layout of this thesis

In Chapter 1 we introduce some terminology concerning Markov chains and finite probability and survey some known results.

In Chapter 2 we examine the relationship between Markov chains and nonnegative matrices. We examine the well-known Perron-Frobenius theorem and its application to the theory of Markov chains.

In Chapter 3 we present the concept of a nearly uncoupled Markov chain and a survey of some of the known properties of such a Markov chain. Algorithmic analysis of nearly uncoupled Markov chains is the focus of this thesis.

We present the concept of the stochastic complement in Chapter 4, along with a number of known theorems concerning its application. The stochastic complement is a tool used to reduce the order of the state space of a given Markov chain. The algorithms we present in this thesis utilise, to great extent, the stochastic complement to produce the nearly uncoupled structure of a given Markov chain.

Chapters 5 and 6 contain our stochastic complement based algorithms, along

with examinations of their properties. The algorithm presented in Chapter 5, the Maximum Entry Algorithm, is our base algorithm. The remainder, presented in Chapter 6, are variations which implement more in-depth reasoning.

In Chapter 7 we conclude our work and present a few notes concerning future research.

Appendices A and B contain extensive calculations involved with solving a very specific stochastic complement related problem. We make use of these calculations in Chapters 5 and 6 to refine the performance of our algorithms. The reader may wish to read these appendices before examining the algorithms in Chapters 5 and 6.

In Appendix C we apply our various algorithms to a number of known Markov chains in order to evaluate and illustrate their performance.

In Appendix D we present sufficient conditions to ensure the success of the Maximum Entry Algorithm when applied to a particularly simple class of Markov chains.

In Appendix E we calculate the complexities (computation time required) of our various algorithms.

Appendix F contains a brief survey of problem matrices – that is, matrices associated with Markov chains which our stochastic complement based algorithms are unlikely to correctly analyse.

Appendix G is a summary of properties of the stochastic complement, which we have included for ease of reference.

# Contents

# Chapter 1

# Finite Probability

## 1.1 Markov chains

A discrete-time *Markov chain* is a *stochastic process* on a finite state space $\mathcal{S}$; it is a sequence of random variables

$$X = \{x_t : t = 0, 1, 2, \ldots\}$$

that take on values in $\mathcal{S}$ and satisfy the Markov property.

We use the notation $\mathbb{P}[a|b]$ to denote the probability that statement $a$ is true, given that statement $b$ is true. As well, $\mathbb{E}[y|b]$ is the expected value of the random variable $y$, given that statement $b$ is true.

**Definition 1.1.** The *Markov property* is the statement that for all $t \geq 0$ and $i_0, \ldots, i_{t-1}, i, j \in \mathcal{S}$,

$$\mathbb{P}\left[x_{t+1} = j | (x_0, x_1, \ldots, x_{t-1}, x_t) = (i_0, i_1, \ldots, i_{t-1}, i)\right] = \mathbb{P}\left[x_{t+1} = j | x_t = i\right].$$

That is, for any $t \geq 0$, the probability distribution of the random variable $x_{t+1}$ is completely determined by the value taken on by $x_t$.

If $x_t = i$ and $x_{t+1} = j$, we say that the Markov chain transitions from $i$ to $j$ at time $t+1$, or that the Markov chain visits state $i$ at time $t$ and state $j$ at time $t+1$.

Such a Markov chain is *time-homogeneous* if for all $i, j \in \mathcal{S}$ and $t_1, t_2 \geq 0$ we have

$$\mathbb{P}\left[x_{t_2+1} = j | x_{t_2} = i\right] = \mathbb{P}\left[x_{t_1+1} = j | x_{t_1} = i\right].$$

In particular, we note that if $X$ is homogeneous in time then, for all $i, j \in \mathcal{S}$ and $t \geq 0$,

$$\mathbb{P}\left[x_{t+1} = j | x_t = i\right] = \mathbb{P}\left[x_1 = j | x_0 = i\right].$$

In other words, the Markov chain is homogeneous in time if, for all $i, j \in \mathcal{S}$, the probability of transitioning from $i$ to $j$ is independent of the time parameter. When $X$ is homogeneous in time, we refer to the value

$$a_{ij} = \mathbb{P}\left[x_{t+1} = j | x_t = i\right] = \mathbb{P}\left[x_1 = j | x_0 = i\right]$$

as the $ij$th transition probability.

The probability distribution of the random variable $x_0$, referred to as the *initial distribution*, is, in general, independent of the transition probabilities and is taken to be given.

We will occasionally make use of the strong Markov property – we first define the notion of a stopping time in order to present the definition of this property.

**Definition 1.2.** Let $X = \{x_t\}$ be a sequence of random variables on a finite state space $\mathcal{S}$. A *stopping time* with respect to $X$ is a random variable $T$ that takes on values in

$$\{0, 1, 2, \ldots\} \cup \{\infty\}$$

such that for any potential value $t'$, the truth (or falsehood) of the statement $T = t'$ is completely determined by the values taken on by $x_0, x_1, \ldots, x_{t'}$. If $t' = \infty$, this refers to the entire sequence.

**Example 1.3.** Let $X$ be a sequence of random variables on a finite state space $\mathcal{S}$; for each $i \in \mathcal{S}$, let

$$T_i = \inf \{t \geq 1 : x_t = i\},$$

with the convention that the *infimum* of the empty set is $\infty$. The case $T_i = 0$ cannot occur, so the falsehood of the statement $T_i = 0$ is trivially determined by $x_0$. For $1 \leq t' < \infty$, $T_i = t'$ if and only if $x_t \neq i$ for all $1 \leq t \leq t' - 1$ and $x_{t'} = i$. Finally, $T_i = \infty$ if and only if $x_t \neq i$ for all $t \geq 1$. Each random variable $T_i$ is a stopping time with respect to $X$.

We will make use of the above stopping time, known as the *first passage time*, in later sections. Stopping times are referred to as such because they often represent the time at which a Markov chain first meets some set condition.

**Definition 1.4.** Let $X = \{x_t\}$ be a sequence of random variables on a state space $\mathcal{S}$. The *strong Markov property* is the statement that if $T$ is a stopping time with respect to $X$, then for all $t \geq 1$ and $i, j \in \mathcal{S}$,

$$\mathbb{P}\left[x_{T+t} = j | T < \infty \text{ and } x_T = i\right] = \mathbb{P}\left[x_t = j | x_0 = i\right].$$

The strong Markov property informs us that if $X = \{x_t\}$ is a Markov chain and $T$ is a stopping time with respect to $X$, then whenever $T \neq \infty$, the sequence

$$X_T = \{x_{T+t} : t = 0, 1, 2, \ldots\}$$

is, itself, a Markov chain with transition probabilities identical to those of $X$ (although, generally, $X_T$ has a different initial distribution than $X$). It can be shown that a sequence of random variables satisfies the strong Markov property if and only if it satisfies the Markov property and is homogeneous in time.

Markov chains with infinite state space, continuous time parameters and/or non-homogeneous transition probabilities are the subject of extensive bodies of research. However, for the course of this work, we will only consider the case of discrete-time, finite state space and time-homogeneous, and simply use the term Markov chain.

## 1.2 Stochastic matrices

We use the notation $\mathbb{1}$ to refer to the column vector with every entry equal to 1 and $e_i$ to refer to the column vector with $i$th entry equal to 1 and every other entry

equal to 0 (the context in which they appear determines the orders of $\mathbb{1}$ and $e_i$).

In general, we will assume that any finite state space under consideration is a subset of $\{1, \ldots, n\}$, for some positive integer $n$.

Let $x$ be a random variable that takes on values in a finite state space $\mathcal{S}$. The column vector $v$ that has

$$v_i = \mathbb{P}[x = i]$$

is referred to as the probability distribution of $x$. Clearly, if $v$ is a probability distribution, every $v_i$ is nonnegative and $v^T \mathbb{1} = 1$ (the sum of the entries in a probability distribution is 1).

Let $X$ be a Markov chain on a finite state space $\mathcal{S}$ and let $v$ be a probability distribution on $\mathcal{S}$. We use the notation $\mathbb{P}_v[*]$ and $\mathbb{E}_v[*]$ to represent the probability measure and the expected value function, respectively, given $x_0$ distributed via $v$. If $v = e_i$, we further abbreviate

$$\mathbb{P}_i[*] = \mathbb{P}_{e_i}[*] \text{ and } \mathbb{E}_i[*] = \mathbb{E}_{e_i}[*].$$

Let $X$ be a Markov chain on the state space $\mathcal{S} = \{1, \ldots, n\}$ and let $A$ be the *matrix of transition probabilities* associated with $X$ (also referred to as the transition matrix of $X$):

$$a_{ij} = \mathbb{P}[x_{t+1} = j | x_t = i] = \mathbb{P}_i[x_1 = j].$$

The $i$th row of $A$ is the probability distribution of the random variable $x_1$ given initial distribution $e_i$. Thus, $A$ is a *stochastic matrix*: it is square, entrywise nonnegative and has the sum of the entries in each row equal to 1. These facts are summarised with the notation

$$A \geq 0 \text{ and } A\mathbb{1} = \mathbb{1}.$$

The following proposition is a well-known fact concerning the transition matrix of a Markov chain (see, for example, [2, Lemma 8.1.2]).

**Proposition 1.5.** *Let $X$ be a Markov chain with transition matrix $A$ and let $v$ be a probability distribution on $\mathcal{S}$, given in vector form. Then, for any $t \geq 1$, $A^t$ is, itself, a stochastic matrix; moreover,*

$$\left(A^t\right)_{ij} = \mathbb{P}_i\left[x_t = j\right] \text{ and } \left(v^T A^t\right)_j = \mathbb{P}_v\left[x_t = j\right].$$

In other words, if $v$ is the probability distribution of $x_0$ then $(v^T A^t)^T$ is the probability distribution of $x_t$.

## 1.3 The transition graph

A *directed graph (digraph)* is an ordered pair $G = (V, E)$ where $V$ is a set, referred to as the *vertices* of $G$, and $E$ is some subset of $V \times V$, referred to as the *directed arcs* of $G$; let $G = (V, E)$ be a digraph.

We label a directed arc $(i, j) \in E$ with the notation $i \to j$ and refer to $i$ and $j$ as the *endpoints,* $i$ as the *initial vertex* and $j$ as the *terminal vertex* of the directed arc.

A *directed walk* of length $t$ in $G$ is a sequence of $t+1$ vertices in $V$ and $t$ directed arcs in $E$ of the form:

$$\omega = i_0 \to i_1 \to \cdots \to i_{t-1} \to i_t.$$

When a directed walk $\omega$ is expressed in this manner, we refer to $i_0$ as the initial vertex and $i_t$ as the terminal vertex of $\omega$ and say that $\omega$ is a directed walk from $i_0$ to $i_t$. We use the notation $\omega : i_0 \rightsquigarrow i_t$ to denote that $\omega$ is such a walk; we use the notation $i \prec_G j$ to represent that $G$ contains a directed walk from $i$ to $j$ with length greater than or equal to 1. Directed walks may have length 0 – these walks consist of a single vertex and no directed arcs. However, if the only directed walk from $i$ to $i$ is the walk $\omega = i$ of length 0, we do not have $i \prec_G i$. The notation $i \preceq_G j$ is used to represent the fact that $j = i$ or $i \prec_G j$.

A directed walk with its initial and terminal vertices identical is a *closed walk,* a directed walk with no repeated vertices is referred to as *directed path* and a directed walk with its initial and terminal vertices equal and no other repeated vertices is a *directed cycle.*

For each nonempty subset $\mathcal{C} \subseteq V$, we define the *induced subgraph* of $G$ corresponding to $\mathcal{C}$ to be the digraph $G(\mathcal{C}) = (\mathcal{C}, E(\mathcal{C}))$ where

$$E(\mathcal{C}) = \{i \to j \in E : i, j \in \mathcal{C}\}.$$

An *isolated vertex* is a vertex that is not an endpoint of any directed arc of $G$. The digraph $G$ is *strongly connected* if for any two vertices $i$ and $j$ we have $i \prec_G j$. A digraph is referred to as *irreducible* if it is strongly connected or consists of a single isolated vertex.

An *irreducible component* of the digraph $G$ is a maximal irreducible induced subgraph. That is, an irreducible component is an induced subgraph $G(\mathcal{C})$ such that either $\mathcal{C} = \{i\}$ where $i$ is isolated, or

1. $G(\mathcal{C})$ is strongly connected, and

2. if $\mathcal{C} \subseteq \mathcal{C}'$ and $G(\mathcal{C}')$ is strongly connected, then $\mathcal{C}' = \mathcal{C}$.

Let $\mathcal{S}$ be a set. A *partition* of $\mathcal{S}$ is a collection of nonempty disjoint subsets of $\mathcal{S}$, $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ such that

$$\mathcal{S} = \bigcup_{k=1}^{m} \mathcal{C}_k.$$

The graph $G = (V, E)$ is *weakly connected* if it consists of a single isolated vertex, or if for any partition of $V$ into two subsets $\mathcal{C}_1$ and $\mathcal{C}_2$ there is at least one directed arc in $E$ with its intial and terminal vertices not contained in the same member of $\{\mathcal{C}_1, \mathcal{C}_2\}$. A *weakly connected component* of $G$ is a maximal weakly connected induced subgraph of $G$. *i.e.* A weakly connected component is an induced subgraph $G(\mathcal{C})$ such that either $\mathcal{C} = \{i\}$ where $i$ is an isolated vertex or

1. for any partition $\{\mathcal{C}_1, \mathcal{C}_2\}$ of $\mathcal{C}$, there is a directed arc in $E$ with one endpoint in $\mathcal{C}_1$ and the other endpoint in $\mathcal{C}_2$, and

2. any directed arc in $E$ that has one endpoint in $\mathcal{C}$ has both endpoints in $\mathcal{C}$.

Let $A$ be a stochastic matrix with state space $\mathcal{S}$. The *transition graph associated with* $A$ is the digraph $G$ with vertex set $V = \mathcal{S}$ and directed arcs

$$E = \{i \to j : a_{ij} \neq 0\}.$$

Let $X$ be a Markov chain with transition matrix $A$; the transition graph of $X$ is simply the transition graph of $A$.

Several probabilistic properties of a Markov chain correspond to combinatorial properties of the associated digraph $G$. For instance, note that $i \prec_G j$ if and only if there is a nonzero probability of transitioning from $i$ to $j$ in a finite number of steps. We will occasionally make use of the terminology $i \prec j$ without explicitly referring to the transition graph of a Markov chain. In this case, we will take $i \prec j$ to refer to the fact that it is possible (there is a nonzero probability) that $x_t = j$ for some $t \geq 1$, given $x_0 = i$.

## 1.4 Essential classes of states

Let $G = (V, E)$ be a digraph. An *essential component* in $G$ is an irreducible component $G(\mathcal{C})$ such that for any arc $i \to j$ in $E$, if $i \in \mathcal{C}$ then $j \in \mathcal{C}$; *i.e.* an

essential component is an irreducible component $G(\mathcal{C})$ that cannot be escaped along a directed arc (or walk). Thus if $G(\mathcal{C})$ is an essential component, $i \in \mathcal{C}$ and $j \notin \mathcal{C}$, then $i \not\to_G j$.

Let $X = \{x_t\}$ be a Markov chain on state space $\mathcal{S}$. An *essential class of states* is a subcollection $\mathcal{E} \subseteq \mathcal{S}$ that forms an essential component in the transition graph of $X$.

For each state $i \in \mathcal{S}$, let $T_i$ be the first passage time:

$$T_i = \inf \{t \geq 1 : x_t = i\}.$$

Further, for each state $i \in \mathcal{S}$ and time $t \geq 1$, let $N_i(t)$ be the random variable that counts the number of passages into state $i$ up to and including time $t$:

$$N_i(t) = |\{s : 1 \leq s \leq t \text{ and } x_s = i\}|.$$

As well, we define

$$N_i = N_i(\infty) = |\{s \geq 1 : x_s = i\}|$$

to be the total number of transitions into $i$. We note that we may very well have $N_i = \infty$.

**Definition 1.6.** Let $X = \{x_t\}$ be a Markov chain on the state space $\mathcal{S}$ and let $i \in \mathcal{S}$. If

$$\mathbb{P}_i [T_i = \infty] = 0,$$

we refer to $i$ as *recurrent*. If state $i$ is not recurrent, it is *transient*.

That is, state $i$ is recurrent if the probability of visiting $i$ exactly once is 0.

Theorems 1.7 and 1.8 summarise known facts concerning recurrent and transient states. See [23, Chapter 1] for a more in-depth examination of these concepts.

**Theorem 1.7.** *Let $X = \{x_t\}$ be a Markov chain on the finite state space $\mathcal{S}$. For each $i \in \mathcal{S}$, the following are equivalent:*

1. *The state $i \in \mathcal{S}$ is recurrent.*
2. *The state $i$ is contained in an essential class of states.*
3. $\mathbb{E}_i[T_i] < \infty$.

**Proof** Let $A$ be the stochastic matrix and $G$ be the transition graph associated with $X$.

Let $i \in \mathcal{S}$ be recurrent and let $\mathcal{E} = \{j : i \prec j\}$. Since $i$ is recurrent, the probability of transitioning from $i$ to $i$ (in one or more steps) is nonzero and so $i \in \mathcal{E}$. Let $j \in \mathcal{E}$ and suppose that $j \not\prec i$; thus, $j \neq i$. As it is impossible to transition from $j$ to $i$, if $T_j < T_i$, then $T_i = \infty$. Let $\omega$ be a directed path $i \rightsquigarrow j$ in $G$ and let

$$\alpha = \prod_{k \to l \in \omega} a_{kl} > 0$$

be the probability of transitioning from $i$ to $j$ along $\omega$. Note that since $\omega$ is a directed path it does not visit any state more than once. Thus, if the Markov chain begins at

$i$ and then transitions into $j$ along $\omega$ we have $T_j < T_i = \infty$ (since $j \not\prec i$, the Markov chain cannot transition from $j$ to $i$). So, we see that

$$\mathbb{P}_i\left[T_i = \infty\right] \geq \mathbb{P}_i\left[T_j < T_i \text{ and } T_i = \infty\right] \geq \alpha > 0;$$

which contradicts the fact that $i$ is recurrent. Thus, if $i \prec j$ then $j \prec i$. Let $j, k \in \mathcal{E}$; then, $j \prec i \prec k$ and so $\mathcal{E}$ is strongly connected. Further, if $j \in \mathcal{E}$ and $j \to k$ is an arc in $G$, then $i \prec j \prec k$ and so $k \in \mathcal{E}$. Thus, $\mathcal{E}$ is an essential class of states.

Now, suppose that $\mathcal{E}$ is an essential class of states. If $\mathcal{E}$ contains only one element, $i$, then we must have $a_{ii} = 1$. This implies that $\mathbb{E}_i[T_i] = 1 < \infty$. So, we assume that $\mathcal{E}$ contains two or more states. Fix a state $i \in \mathcal{E}$; for each $j \in \mathcal{E}$ distinct from $i$, let

$$\beta(j) = \mathbb{E}_i\left[N_j(T_i)\right] = \sum_{m \geq 1} m\mathbb{P}_i\left[N_j(T_i) = m\right]$$

be the expected number of transitions into $j$ between two visits to $i$.

We note that if

$$\mathbb{P}_i\left[N_j(T_i) = \infty\right] > 0,$$

then $\beta(j) = \infty$.

As $\mathcal{E}$ forms an essential class, for any $j \in \mathcal{E} \setminus i$, there is a directed path $\omega_j : j \rightsquigarrow i$ in $G$. Let

$$\alpha(j) = \prod_{(k \to l) \in \omega_j} a_{kl} > 0$$

be the probability of transitioning from $j$ to $i$ along $\omega_j$. Thus, after visiting state $j$ the probability that the Markov chain visits $i$ before another visit to $j$ is at least $\alpha(j)$.

First we show that

$$\mathbb{P}_i\left[N_j(T_i) = \infty\right] = 0.$$

Every time the Markov chain visits $j$, the probability that it then visits $j$ again before a visit to $i$ is less than $1 - \alpha(j)$. Thus,

$$\mathbb{P}_i\left[N_j(T_i) \geq m\right] \leq (1 - \alpha(j))^m,$$

for all positive integers $m$. Taking the limit as $m \to \infty$ show the above claim.

So, we have

$$\beta(j) = \mathbb{E}_i\left[N_j(T_i)\right] = \sum_{m \geq 1} m\mathbb{P}_i\left[N_j(T_i) = m\right] = \mathbb{E}_i\left[N_j(T_i)\right] = \sum_{1 \leq m < \infty} m\mathbb{P}_i\left[N_j(T_i) = m\right].$$

Then, for all $m \geq 0$,

$$\begin{aligned} \mathbb{P}_i\left[N_j(T_i) = m+1\right] &\leq \mathbb{P}_i\left[N_j(T_i) \geq m+1\right] \\ &\leq (1 - \alpha(j))\,\mathbb{P}_i\left[N_j(T_i) = m\right]. \end{aligned}$$

This implies that for all positive integers $m$,

$$\mathbb{P}_i\left[N_j(T_i) = m\right] \leq (1 - \alpha(j))^{m-1}\,\mathbb{P}_i\left[N_j(T_i) = 1\right].$$

We further note that

$$\mathbb{P}_i[N_j(T_i) = 1] \;\leq\; \mathbb{P}_i[N_j(T_i) \geq 1] \;=\; \mathbb{P}_i[T_j < T_i].$$

Thus,

$$\begin{aligned}
\beta(j) &= \sum_{m=1}^{\infty} m\mathbb{P}_i\left[N_j(T_i) = m\right] \\
&\leq \sum_{m=1}^{\infty} m\left(1-\alpha(j)\right)^{m-1}\mathbb{P}_i\left[T_j < T_i\right] \\
&= \mathbb{P}_i\left[T_j < T_i\right]/\alpha(j)^2.
\end{aligned}$$

And so we conclude that for any $j \in \mathcal{E}$ with $j \neq i$, $\beta(j) = \mathbb{E}_i[N_j(T_i)]$ is not equal to $\infty$. Now, if $x_0 \in \mathcal{E}$, then

$$T_i = 1 + \sum_{j \in \mathcal{E}\setminus i} N_j(T_i).$$

This shows that

$$\mathbb{E}_i\left[T_i\right] = 1 + \sum_{j \in \mathcal{E}\setminus i} \beta(j) < \infty.$$

Finally, suppose that $\mathbb{E}_i[T_i] < \infty$. It is clear that, given $x_0 = i$, the probability that $T_i = \infty$ must be 0; state $i$ is recurrent.

Thus, we have shown that condition 1 implies condition 2, 2 implies 3 and 3 implies 1. ∎

**Theorem 1.8.** *Let $X = \{x_t\}$ be a Markov chain on a finite state space $\mathcal{S}$. Then,*

1. *$\mathcal{S}$ contains an essential class of states;*

2. *for any state $i$ there is a recurrent state $j$ such that $i \prec j$; and*

3. *for any state $i$, if there is a recurrent state $j$ such that $j \prec i$, then $i$ is recurrent.*

**Proof** We use a proof by induction on the number of states in $\mathcal{S}$. If $\mathcal{S}$ contains only one element, all three statements are trivial. So, suppose that $|\mathcal{S}| \geq 2$ and that the first statement holds for all such Markov chains on state spaces containing strictly fewer than $|\mathcal{S}|$ states.

Suppose that $i \prec j$ for all pairs of states in $\mathcal{S}$. This implies that the transition graph of $X$ is strongly connected – the entire state space $\mathcal{S}$ is a single essential class of states and, by Theorem 1.7, every state is recurrent.

So, assume that $i \not\prec j$ for some $i, j \in \mathcal{S}$ (possibly $i = j$). Let

$$\mathcal{S}_i = \{k : i \prec k\}$$

and let $T$ be the stopping time

$$T = \inf \{t \geq 0 : x_t \in \mathcal{S}_i\}.$$

Note that if $k \in \mathcal{S}_i$ and $l \notin \mathcal{S}_i$, then $k \not\prec l$. Thus, for $t \geq T$ (if $T < \infty$), $x_t \in \mathcal{S}_i$. So, we use the strong Markov property to define the Markov chain $Y = \{y_t\} = \{x_{T+t}\}$ on state space $\mathcal{S}_i$. Since $j \notin \mathcal{S}_i$, $\mathcal{S}_i$ contains strictly fewer states than $\mathcal{S}$. By the inductive hypothesis, it contains an essential class $\mathcal{E}$, with respect to $Y$. An essential

class with respect to $Y$ is an essential class with respect to $X$, and so $\mathcal{S}$ contains an essential class of states.

Next, consider an arbitrary state $i \in \mathcal{S}$. First suppose that there is $j \in \mathcal{S}$ with $i \not\prec j$. The above reasoning shows that $\mathcal{S}_i = \{k : i \prec k\}$ contains an essential class $\mathcal{E}$ and so $\mathcal{S}_i$ contains a recurrent state $k$, which then has $i \prec k$. If we suppose that $i \prec j$ for every $j \in \mathcal{S}$, then $i \prec k$ for some recurrent $k$, since statement 1 implies that $\mathcal{S}$ contains at least one recurrent state.

To see that the final statement holds, assume that $j$ is recurrent and that $j \prec i$. Then, $j$ is contained in an essential class of states, $i$ must be contained in that same essential class and so $i$ is, itself, recurrent (via Theorem 1.7). ∎

# Chapter 2

# Stochastic matrices and eigenvalues of Markov chains

## 2.1 Definitions

Let $A$ be a square matrix with index set $\mathcal{S}$. When $A$ is a stochastic matrix, we refer to $\mathcal{S}$ as the state space of $A$ and the elements of $\mathcal{S}$ as its states. Let $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathcal{S}$. We define the $(\mathcal{C}_1, \mathcal{C}_2)$-submatrix of $A$ to be the matrix

$$A(\mathcal{C}_1, \mathcal{C}_2) = [a_{ij}]_{i \in \mathcal{C}_1, j \in \mathcal{C}_2} .$$

When $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$, we refer to this submatrix as the *principal submatrix* of $A$ corresponding to $\mathcal{C}$, and use the abbreviation $A(\mathcal{C}) = A(\mathcal{C}, \mathcal{C})$. We will occasionally refer to principal submatrices of $A$ without reference to any collections $\mathcal{C}$ – these are simply principal submatrices corresponding to some collection of indices.

Similarly, if $v$ is a column vector on the index set $\mathcal{S}$ and $\mathcal{C} \subseteq \mathcal{S}$, the subvector corresponding to $\mathcal{C}$ is the column vector

$$v(\mathcal{C}) = [v_i]_{i \in \mathcal{C}} .$$

When we consider submatrices (or subvectors) determined by collections $\mathcal{C}$, we will use the same indices $\mathcal{C}$ to reference the entries of the submatrix (or subvector). For example, let $A$ be a square matrix on the indices $\mathcal{S} = \{1, 2, 3, 4, 5\}$, let $\mathcal{C} = \{1, 3, 4\}$ and let $B = A(\mathcal{C})$; then,

$$B = \begin{bmatrix} b_{11} & b_{13} & b_{14} \\ b_{31} & b_{33} & b_{34} \\ b_{41} & b_{43} & b_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{13} & a_{14} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{bmatrix} .$$

For example, the $(3, 3)$th entry of $B = A(\mathcal{C})$ is $b_{33} = a_{33}$ (not $a_{44}$). If $i \notin \mathcal{C}_1$ or $j \notin \mathcal{C}_2$, then there is no $(i, j)$th entry of $A(\mathcal{C}_1, \mathcal{C}_2)$, even if $i$ and $j$ are integers smaller, respectively, than the orders of $\mathcal{C}_1$ and $\mathcal{C}_2$.

Let $A$ be a square matrix with associated digraph $G$. If $G$ is irreducible, we refer to $A$ as irreducible. An *irreducible block* of $A$ is a principal submatrix $A(\mathcal{C})$ where $G(\mathcal{C})$ is an irreducible component of $G$. We refer to a Markov chain as irreducible if its transition digraph and matrix are irreducible.

A *permutation matrix* is a matrix $P$ that has every entry equal to 0 or 1 and has exactly one entry equal to 1 in every row and column. When $P$ is a permutation matrix, $P$ is nonsingular and $P^T = P^{-1}$.

Let $A$ and $B$ be square matrices with index sets $\mathcal{S}_A$ and $\mathcal{S}_B$. We say that the matrices $A$ and $B$ are *permutation-similar* if there is a permutation matrix $P$ such that $B = PAP^T$; we use the notation $A \cong B$ to represent this fact.

It is straightforward to show that $A$ and $B$ are permutation-similar if and only if there is a bijection $f : \mathcal{S}_B \mapsto \mathcal{S}_A$ such that $b_{ij} = a_{f(i)f(j)}$ for all $i$ and $j$. When this holds, the permutation matrix $P$ that accomplishes the similarity is the matrix whose rows are indexed by $\mathcal{S}_B$, whose columns are indexed by $\mathcal{S}_A$ and has

$$p_{ij} = \begin{cases} 1 & \text{if } j = f(i), \\ 0 & \text{if } j \neq f(i). \end{cases}$$

The following lemma is a standard result in combinatorial matrix theory (see [3, Chapter 3], for example) and will aid us in our examinations of stochastic matrices.

**Lemma 2.1.** *Let $A$ be a square matrix on index set $\mathcal{S}$ and let $G$ be the digraph associated with $A$. Then, $A$ is permutation-similar to a matrix of the form*

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ * & & A_m \end{bmatrix},$$

*where the matrices $A_k$ are the irreducible blocks of $A$. Moreover, this expression is unique, up to reordering the indices within each block and possibly reordering the diagonal blocks.*

Let $X$ be a Markov chain with state space $\mathcal{S}$, transition matrix $A$ and transition graph $G$. The above lemma implies the existence of one-to-one correspondences between each of

1. the irreducible blocks of $A$;

2. the irreducible components of $G$,

3. and a partition of the state space $\mathcal{S}$.

Further, there is a one-to-one correspondence between subcollections of each of these and the collection of the essential classes of states with respect to $X$.

## 2.2 The Perron-Frobenius theorem

We present the Perron-Frobenius Theorem, as it applies to stochastic matrices. Various versions of this well-known theorem (and their proofs) can be found in [3, 12, 14].

Let $A$ be a square matrix with spectrum (eigenvalues) $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$; the *spectral radius* of $A$ is the maximum modulus among the eigenvalues $\sigma(A)$:

$$\rho(A) = \max_{\lambda \in \sigma(A)} \{|\lambda|\}.$$

A nonnegative matrix or vector is one where every entry is a nonnegative real number; a positive matrix or vector is one where every entry is a positive number.

Let $A$ and $B$ be matrices, or vectors, of the same order. If $B - A$ is nonnegative, that is, if $b_{ij} \geq a_{ij}$ for all $i$ and $j$, we say that $B \geq A$. Similarly, if $B - A$ is positive we say that $B > A$. We use the notation $B \neq A$ to signify the fact that for at least one pair $i$ and $j$, $b_{ij} \neq a_{ij}$. (In particular, the notation $A \neq 0$ represents the statement that $A$ has at least one nonzero entry, not that every entry of $A$ is nonzero.)

An algebraically simple eigenvalue of a square matrix is an eigenvalue whose multiplicity as a root of the characteristic polynomial is 1. Thus, the generalised eigenspace and the eigenspace associated with an algebraically simple eigenvalue are identical and have dimension equal to 1. See [14] for a full discussion on the distinction between eigenspaces and generalised eigenspaces.

**Perron-Frobenius Theorem.** *Let $A$ be an irreducible nonnegative square matrix and suppose that $A \neq [0]$. Then,*

1. *$\rho(A) > 0$;*
2. *$\rho(A)$ is an algebraically simple eigenvalue of $A$;*
3. *the eigenspace associated with $\rho(A)$ is spanned by a positive eigenvector;*
4. *if $Av = \lambda v$ where $v \neq 0$ and $v \geq 0$, then $\lambda = \rho(A)$ and, in fact, $v > 0$; and*
5. *if $B$ is a nonnegative matrix such that $B \leq A$ and $B \neq A$, then $\rho(B) < \rho(A)$.*

Let $A$ be irreducible and nonnegative. The positive number $\rho(A)$ is referred to as the *Perron value* of $A$; the unique positive vector $v$ that satisfies

$$Av = \rho(A)v \text{ and } \mathbb{1}^T v = 1$$

is referred to as the *Perron vector* of $A$. The Perron vector of $A^T$ is referred to as the *left Perron vector* of $A$, as we have

$$w^T A = \rho(A) w^T, \text{ whenever } A^T w = \rho(A) w.$$

Theorem 2.2 appears in [21, Theorem 1.7.5].

**Theorem 2.2.** *Let $A$ be an irreducible stochastic matrix. The Perron value of $A$ is 1. Moreover, let $\pi$ be the left Perron vector of $A$; then,*

$$\frac{\pi(j)}{\pi(i)} = \mathbb{E}_i \left[ N_j(T_i) \right] \text{ and } \frac{1}{\pi(i)} = \mathbb{E}_i \left[ T_i \right],$$

*with respect to the Markov chain associated with $A$.*

**Remark.** That is, $\pi(j)/\pi(i)$ is the expected number of visits to state $j$ between two visits to state $i$, and $1/\pi(i)$ is the expected amount of time between two visits to state $i$. We refer to the left Perron vector $\pi$ of an irreducible stochastic matrix as its *stationary distribution*.

**Proof** Since $A\mathbb{1} = \mathbb{1}$, the Perron-Frobenius theorem implies that $\rho(A) = 1$. Let $\pi$ be the stationary distribution (left Perron vector) of $A$. Fix a state $i \in \mathcal{S}$ and let $v$ be the vector with

$$v_j = \mathbb{E}_i\left[N_j(T_i)\right] = \sum_{t\geq 1} \mathbb{P}_i\left[x_t = j \text{ and } t \leq T_i\right].$$

We note that if $j \neq i$, then

$$v_j = \sum_{t\geq 1} \mathbb{P}_i\left[x_t = j \text{ and } t \leq T_i\right] = \sum_{t\geq 1} \mathbb{P}_i\left[x_t = j \text{ and } t+1 \leq T_i\right]$$

(since $x_{T_i} = i$). When $T_i < \infty$, $N_i(T_i) = 1$ with probability 1; and so, since $i$ is recurrent, $v_i = 1$. In the proof of Theorem 1.7, we saw that each $v_j$ is finite ($v_j = \beta(j)$, in that proof). Now, for all $j$,

$$\begin{aligned}
\left(v^T A\right)_j &= \sum_k v_k a_{kj} \\
&= a_{ij} + \sum_{k\neq i} v_k a_{kj} \\
&= \mathbb{P}_i\left[x_1 = j\right] + \sum_{k\neq i}\sum_{t\geq 1} \mathbb{P}_i\left[x_t = k \text{ and } t+1 \leq T_i\right] \mathbb{P}\left[x_{t+1} = j | x_t = k\right].
\end{aligned}$$

Now, we note that

$$\mathbb{P}_i\left[x_t = k \text{ and } t+1 \leq T_i\right] \mathbb{P}\left[x_{t+1} = j | x_t = k\right]$$

$$= \mathbb{P}_i\left[x_t = k \text{ and } i \notin \{x_1, \ldots, x_{t-1}\}\right] \mathbb{P}\left[x_{t+1} = j | x_t = k\right]$$

$$= \mathbb{P}_i\left[x_t = k,\ x_{t+1} = j \text{ and } i \notin \{x_1, \ldots, x_{t-1}\}\right]$$

$$= \mathbb{P}_i\left[x_t = k,\ x_{t+1} = j \text{ and } t+2 \leq T_i\right].$$

So,

$$\begin{aligned}
\left(v^T A\right)_j &= \mathbb{P}_i\left[x_1 = j\right] + \sum_{k \neq i} \sum_{t \geq 1} \mathbb{P}_i\left[x_t = k,\ x_{t+1} = j \text{ and } t+2 \leq T_i\right] \\
&= \mathbb{P}_i\left[x_1 = j\right] + \sum_{t \geq 1} \sum_{k \neq i} \mathbb{P}_i\left[x_t = k,\ x_{t+1} = j \text{ and } t+2 \leq T_i\right] \\
&= \mathbb{P}_i\left[x_1 = j\right] + \sum_{t \geq 2} \mathbb{P}_i\left[x_t = j \text{ and } t+1 \leq T_i\right] \\
&= \sum_{t \geq 1} \mathbb{P}_i\left[x_t = j \text{ and } t+1 \leq T_i\right] \\
&= \mathbb{E}_i\left[N_j(T_i)\right] \\
&= v_j.
\end{aligned}$$

Thus, $v^T A = v^T$. By the Perron-Frobenius theorem, $v$ is a scalar multiple of $\pi$. Since $v_i = 1$, we must have

$$v = \frac{1}{\pi_i}\pi.$$

Therefore,

$$\mathbb{E}_i\left[N_j(T_i)\right] = v_j = \frac{\pi_j}{\pi_i}.$$

The first equality holds. Finally, we note that

$$1 + \sum_{j \neq i} N_j(T_i) = T_i \text{ and } \sum_{j \neq i} \pi_j = 1 - \pi_j.$$

This implies that

$$\begin{aligned}
\mathbb{E}_i\left[T_i\right] &= 1 + \sum_{j \neq i} \mathbb{E}_i\left[N_j(T_i)\right] \\
&= 1 + \sum_{j \neq i} \tfrac{\pi_j}{\pi_i} \\
&= 1 + \tfrac{1-\pi_i}{\pi_i} \\
&= \tfrac{1}{\pi_i},
\end{aligned}$$

proving the second equality. ∎

Theorem 2.3 follows from the Perron Frobenius theorem together with the material contained in [23, Chapter 1].

**Theorem 2.3.** *Let $A$ be a stochastic matrix. Then, $\rho(A) = 1$; moreover, the multiplicity of* 1 *as an eigenvalue of $A$ is equal to the number of essential classes contained in the state space of the Markov chain associated with $A$.*

**Proof** By Lemma 2.1, we can express $A$ as

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ * & & A_m \end{bmatrix},$$

where each $A_k$ is an irreducible square matrix.

Suppose that $A_k$ corresponds to an essential class of states. The probability of transitioning out of an essential class is 0; so the entries contained in the same rows as $A_k$ but outside of $A_k$ must all be 0. The matrix $A_k$ is an irreducible stochastic matrix and so has Perron value equal to 1.

Suppose that $A_k$ does not correspond to an essential class; let the states associated with $A_k$ be $\mathcal{C}_k$. Since $A_k$ is irreducible and $\mathcal{C}_k$ is not essential, there must be $i \in \mathcal{C}_k$ and $j \notin \mathcal{C}_k$ such that $a_{ij} > 0$. Thus, at least one row of $A_k$ has sum strictly less than 1. Let $D_k$ be the diagonal matrix whose diagonal entries are the corresponding entries of

$$\mathbb{1} - A_k\mathbb{1}$$

and let $A'_k = A_k + D_k$. The matrix $A'_k$ is then irreducible and stochastic and so $\rho(A'_k) = 1$. Then, we note that $A'_k - A_k = D_k \neq 0$ is nonnegative and so $\rho(A_k) < 1$.

Thus, since the eigenvalues of $A$ are those of its diagonal blocks, the multiplicity of 1 as an eigenvalue of $A$ is equal to the number of essential classes. We further note that the associated state space must contain at least one essential class of states and so $\rho(A) = 1$. ∎

**Corollary 2.4.** *Let $A$ be a stochastic matrix and let $A_1, \ldots, A_m$ be the principal submatrices of $A$ corresponding to essential classes. The left eigenspace of $A$ corresponding to 1 has an orthogonal basis consisting of vectors of the form*

$$\pi_k^T \cong \begin{bmatrix} \hat{\pi}_k^T & 0^T \end{bmatrix},$$

*where $\hat{\pi}_k$ is the stationary distribution of $A_k$ and the support of $\pi_k$ consists of the states corresponding to $A_k$.*

Let $A$ be a stochastic matrix. A stationary distribution of $A$ is a left eigenvector $\pi$ associated with 1 that has each entry nonnegative and $\pi^T \mathbb{1} = 1$. When $A$ is irreducible there is a unique stationary distribution; however, in general, the stationary distributions are the convex hull of the vectors described in Corollary 2.4.

Let $X = \{x_t\}$ be a Markov chain on state space $\mathcal{S}$ and let $\mathcal{C} \subseteq \mathcal{S}$ be a collection of states. We say that the random process $X$ enters $\mathcal{C}$ if there is some $t \geq 1$ with $x_t \in \mathcal{C}$. (If every $x_t \in \mathcal{C}$ we still say that $X$ enters $\mathcal{C}$.)

Theorem 2.5 is an expanded version of [23, Theorem 4.4].

**Theorem 2.5.** *Let $X$ be a Markov chain with state space $\mathcal{S}$ and transition matrix $A$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the essential classes of states contained in $\mathcal{S}$; for each $i \in \mathcal{S}$ and $1 \leq k \leq m$, let*

$$v_k(i) = \mathbb{P}_i \left[ X \text{ enters } \mathcal{E}_k \right]$$

*be the probability that, given $x_0 = i$, the Markov chain will enter $\mathcal{E}_k$. Then, the vectors $v_1, \ldots, v_m$ form a basis of the right eigenspace of $A$ associated with $\rho(A) = 1$.*

**Proof** Note that if $x_0 \in \mathcal{E}_l$ then $x_t \in \mathcal{E}_l$ for all $t \geq 1$; and so

$$v_k(\mathcal{E}_l) = \begin{cases} \mathbb{1} & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

This implies that the $m$ vectors $v_k$ are linearly independent. Since $m$ is the multiplicity of 1 as an eigenvalue, we simply need to show that $Av_k = v_k$. Fix a state $i \in \mathcal{S}$ and consider the Markov chain with initial distribution $e_i$. Since each $\mathcal{E}_k$ is essential, if $x_t \in \mathcal{E}_k$ then $x_s \in \mathcal{E}_k$ for all $s \geq t$. Thus, $\{x_0, x_1, \ldots\}$ enters $\mathcal{E}_k$ if and only if $\{x_1, x_2, \ldots\}$ enters $\mathcal{E}_k$. As $x_0 = i$, the probability distribution of $x_1$ is

$$r_i^T = \begin{bmatrix} a_{i1} & \cdots & a_{in} \end{bmatrix}.$$

So,

$$
\begin{aligned}
v_k(i) &= \mathbb{P}_i\left[X \text{ enters } \mathcal{E}_k\right] \\
&= \mathbb{P}_{r_i}\left[X \text{ enters } \mathcal{E}_k\right] \\
&= \sum_j \mathbb{P}_{r_i}\left[x_0 = j \text{ and } X \text{ enters } \mathcal{E}_k\right] \\
&= \sum_j a_{ij}\mathbb{P}_j\left[X \text{ enters } \mathcal{E}_k\right] \\
&= \sum_j a_{ij} v_k(j) \\
&= \left(Av_k\right)(i).
\end{aligned}
$$

Therefore, $Av_k = v_k$. ∎

## 2.3 Substochastic matrices

A square matrix $A$ is *substochastic* if it is an entrywise nonnegative square matrix and the sum of the entries in each row is less than or equal to 1:

$$A \geq 0 \text{ and } A\mathbb{1} \leq \mathbb{1}.$$

Principal submatrices of stochastic matrices are substochastic. We refer to a matrix

$A$ as *properly substochastic* if it is substochastic and no principal submatrix of $A$ (including $A$ itself) is stochastic.

**Proposition 2.6.** *Let $A$ be a substochastic matrix. Then, $\rho(A) \leq 1$; moreover, $A$ is properly substochastic if and only if $\rho(A) < 1$.*

**Proof** We express $A$ in lower-triangular form,

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ * & & A_m \end{bmatrix},$$

where each $A_k$ is irreducible. So,

$$\rho(A) = \sup\{\rho(A_k)\}.$$

For each $A_k$, let $D_k$ be the unique diagonal matrix such that $A_k\mathbb{1} + D_k\mathbb{1} = \mathbb{1}$. Since $A\mathbb{1} \leq \mathbb{1}$, we have $A_k\mathbb{1} \leq \mathbb{1}$ for each $k$ and so $D_k \geq 0$. Thus, each $A_k + D_k$ is stochastic and so $\rho(A_k) \leq \rho(A_k + D_k) = 1$ with equality if and only if $A_k$ is stochastic ($D_k = 0$) . Thus, $\rho(A) \leq 1$, with equality if and only if one of the irreducible blocks of $A$ is stochastic.

Now, suppose that $A$ is properly substochastic. Then, none of the principal submatrices of $A$, including its irreducible blocks, are stochastic. So, via the above reasoning, $\rho(A) < 1$.

Conversely, assume that $\rho(A) < 1$. Suppose further that $A$ has a principal submatrix that is stochastic. Express

$$A \cong \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11}$ is stochastic. We must then have $A_{12} = 0$, as $A_{11}\mathbb{1} = \mathbb{1}$ and $A_{11}\mathbb{1} + A_{12}\mathbb{1} \leq \mathbb{1}$. Then, $\rho(A) \geq \rho(A_{11}) = 1$, contradicting our initial assumption that $\rho(A) < 1$. So, if $\rho(A) < 1$, $A$ is properly substochastic. ∎

**Corollary 2.7.** *Let $A$ be a stochastic matrix, let $\mathcal{C}$ be a collection of indices of $A$ and let $B = A(\mathcal{C})$ be the principal submatrix of $A$ corresponding to $\mathcal{C}$. Then, $\rho(B) < 1$ if and only if $\mathcal{C}$ does not contain an essential class of states. That is, $B$ is properly substochastic if and only if $\mathcal{C}$ does not contain an essential class of states.*

Let $B$ be a substochastic matrix. We will typically assume that $B$ is a principal submatrix of some stochastic matrix $A$, even if $A$ is not explicitly given; thus, we will refer to the indices of $B$ as its state space. We will define an essential class of states, with respect to $B$, to be some collection of states $\mathcal{E}$ such that the principal submatrix $B(\mathcal{E})$ is irreducible and stochastic. Let $B = A(\mathcal{C})$ be a principal submatrix of a stochastic matrix $A$ associated with a collection of states $\mathcal{C}$. Then, a collection $\mathcal{E} \subseteq \mathcal{C}$ is essential with respect to $B$ if and only if it is essential with respect to $A$, as we then have $B(\mathcal{E}) = A(\mathcal{E})$.

**Lemma 2.8.** *Let $B$ be a substochastic matrix. Then, $I - B$ is nonsingular if and only if $B$ is properly substochastic. Moreover, when $B$ is properly substochastic,*

$$(I - B)^{-1} = \sum_{s=0}^{\infty} B^s.$$

*Thus, $(I - B)^{-1}$ is entrywise nonnegative when it exists.*

**Proof** The matrix $I - B$ is nonsingular if and only if 1 is not an eigenvalue of $B$. Via Corollary 2.7, 1 is not an eigenvalue of the substochastic matrix $B$ if and only if $B$ is properly substochastic.

Let $B$ be properly substochastic. For complex numbers $z$ with $|z| < 1$, we have

$$(1 - z)^{-1} = \sum_{s=0}^{\infty} z^s.$$

Thus, since the eigenvalues $\lambda$ of $B$ satisfy $|\lambda| < 1$, the Neumann series

$$\sum_{s=0}^{\infty} B^s$$

converges to $(I-B)^{-1}$ (see [14, Section 5.6]). Since each $B^s$ is entrywise nonnegative, $(I - B)^{-1}$ is entrywise nonnegative. ∎

## 2.4 Reversible stochastic matrices

A diagonal matrix is a matrix $C$ that has every off-diagonal entry equal to 0 – that is, $c_{ij} = 0$ whenever $i \neq j$. When $C$ is diagonal, we use the shortened notation

$c_i = c_{ii}$. Let $C$ be a diagonal matrix; if $c_i \geq 0$ for all $i$ we refer to $C$ as a *nonnegative diagonal matrix,* and if $c_i > 0$ for all $i$ we refer to $C$ as a *positive diagonal matrix.*

A brief introduction to reversible stochastic matrices and Markov chains, which includes Proposition 2.12, appears in [21, Section 1.9].

**Definition 2.9.** Let $A$ be a stochastic matrix. We say that $A$ is *reversible* if there is a positive diagonal matrix $\Pi$ such that $\Pi A$ is symmetric. A reversible Markov chain is one with a reversible stochastic matrix. A substochastic matrix is reversible under the same conditions as a stochastic matrix; that is, the substochastic matrix $B$ is reversible if there is a positive diagonal matrix $\Pi$ such that $\Pi B$ is symmetric.

**Proposition 2.10.** *A reversible Markov chain has no transient states.*

**Proof** Let $A$ be stochastic and let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Therefore $\pi_i a_{ij} = \pi_j a_{ji}$ for all $i$ and $j$. Each $\pi_i$ is nonzero; so, whenever $a_{ij} \neq 0$ we also have $a_{ji} \neq 0$. This clearly implies that whenever $i \prec j$ we have $j \prec i$. So, let $i$ be a state in the associated state space; by Theorem 1.8, there is a recurrent state $j$ with $i \prec j$. Thus, there is a recurrent state $j$ such that $j \prec i$. The state $i$ must be recurrent. ∎

**Corollary 2.11.** *A reversible stochastic matrix is permutation-similar to a block-diagonal matrix where each block is an irreducible reversible stochastic matrix.*

**Proof** Let $A$ be reversible and let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Since $A$ has no transient states, it is permutation-similar to a block diagonal matrix where each block is irreducible and stochastic. Let $A'$ be an irreducible block of $A$ and let $\Pi'$ be the corresponding principal submatrix of $\Pi$. Then, $\Pi' A'$ is an irreducible principal submatrix of the symmetric matrix $\Pi A$ and is, itself, symmetric. Thus, the irreducible blocks of $A$ are reversible. ∎

**Proposition 2.12.** *Let $A$ be a reversible stochastic matrix and let $\Pi \neq 0$ be a nonnegative diagonal matrix. Then, $\Pi A$ is symmetric if and only if $\Pi \mathbb{1}$ is a left eigenvector of $A$ associated with 1. Thus, if $A$ is reversible and $\pi$ is a stationary distribution of $A$, then for all $i$ and $j$,*

$$\pi_i a_{ij} = \pi_j a_{ji}.$$

**Proof** First, suppose that $A$ is reversible and let $\Pi \neq 0$ be a nonnegative diagonal matrix such that $\Pi A$ is symmetric. Then, for $\pi = \Pi \mathbb{1}$,

$$\pi^T A = \mathbb{1}^T \Pi A = \mathbb{1}^T A^T \Pi$$

$$= \mathbb{1}^T \Pi = \pi^T.$$

Since $\pi = \Pi \mathbb{1}$, $\pi \neq 0$ and we see that $\pi$ is a left eigenvector of $A$ associated with 1.

Now, suppose that $A$ is reversible and that $\Pi \neq 0$ is a nonnegative diagonal matrix such that $\pi = \Pi \mathbb{1}$ is a left eigenvector of $A$ associated with 1. By Corollary 2.11,

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{bmatrix},$$

where each $A_k$ is irreducible and reversible. Let

$$\Pi \cong \begin{bmatrix} \Pi_1 & & 0 \\ & \ddots & \\ 0 & & \Pi_m \end{bmatrix} \text{ and } \pi^T \cong \begin{bmatrix} \nu_1^T & \dots & \nu_m^T \end{bmatrix}$$

be the expressions of $\Pi$ and $\pi$ corresponding to the above expression of $A$. We note that for all $k$, $\nu_k^T A_k = \nu_k^T$. For each $k$, let $\Pi_k'$ be a positive diagonal matrix such that $\Pi_k' A_k$ is symmetric and let $\varpi_k = \Pi_k' \mathbb{1}$.

Let $1 \leq k \leq m$. Via our initial reasoning ($\Pi_k' A_k$ is symmetric), $\varpi_k^T A_k = \varpi_k^T$. Thus, since $A_k$ is irreducible, $\varpi_k > 0$ and $\nu_k \geq 0$, the Perron-Frobenius Theorem implies that there is a nonnegative number $\alpha_k$ such that $\nu_k = \alpha_k \varpi_k$. So, $\Pi_k = \alpha_k \Pi_k'$, further implying that

$$\Pi_k A_k = \alpha_k \Pi_k' A_k$$

is symmetric. The block diagonal structure of $A$ and $\Pi$ then implies that $\Pi A$ is symmetric.

Finally, let $\pi$ be a stationary distribution of $A$ and let $\Pi$ be the diagonal matrix with $i$th diagonal entry equal to $\pi_i$. By the above, $\Pi A$ is symmetric, implying that

for all $i$ and $j$,

$$\pi_i a_{ij} = (\Pi A)_{ij} = \left(A^T \Pi\right)_{ij} = \pi_j a_{ji}.$$

∎

Proposition 2.13 is taken from [7, Proposition 2.2].

**Proposition 2.13.** *Let $A$ be a reversible stochastic matrix. Then, $A$ is diagonally similar to a symmetric matrix. Thus, the eigenvalues of $A$ are real.*

**Proof** Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. We define $\Pi^{1/2}$ and $\Pi^{-1/2}$ to be the positive diagonal matrices whose diagonal entries are $\sqrt{\pi_{ii}}$ and $1/\sqrt{\pi_{ii}}$, respectively. Then,

$$\begin{aligned} \left(\Pi^{1/2} A \Pi^{-1/2}\right)^T &= \Pi^{-1/2} A^T \Pi^{1/2} &&= \Pi^{-1/2} A^T \Pi \Pi^{-1/2} \\ &= \Pi^{-1/2} \Pi A \Pi^{-1/2} &&= \Pi^{1/2} A \Pi^{-1/2}. \end{aligned}$$

∎

### 2.4.1 The reverse of a stochastic matrix

A reversible stochastic matrix has an interesting interpretation as a model of a Markov chain. For some stochastic processes, we can imagine running the process backwards or "rewinding" the Markov chain.

**Definition 2.14.** Let $X$ be an irreducible time-homogeneous Markov chain on a finite state space $\mathcal{S}$ with a discrete time parameter; let $A$ be the associated transition matrix and let $\pi$ be the unique stationary distribution. The *reverse* Markov chain $X^R$ is the random process on the state space $\mathcal{S}$ that has its $ij$th transition probability equal to

$$a_{ij}^R = \frac{\pi_j a_{ji}}{\pi_i}.$$

We provide exposition showing that the reverse of a Markov chain can be seen as a way of reversing the process in time. Let $X = \{x_t\}$ be an irreducible Markov chain on the state space $\mathcal{S}$. Let $A$ be the associated transition matrix and let $\pi$ be the stationary distribution of $A$; suppose further that the initial distribution of $X$ is $\pi$. Proposition 1.5, together with the fact that $\pi^T A = \pi^T$, implies that

$$\mathbb{P}_\pi [x_t = i] = \left(\pi^T A^t\right)_i = \pi_i.$$

Let $Y = \{y_t\}$ be the Markov chain on $\mathcal{S}$ whose $ij$th transition probability is the probability that if $X$ has transitioned into $i$, the preceding state was $j$:

$$b_{ij} = \mathbb{P}_i [y_1 = j] = \mathbb{P}_\pi [x_s = j | x_{s+1} = i]$$

Now,

$$\begin{aligned}
\mathbb{P}_\pi\left[x_s = j | x_{s+1} = i\right] &= \frac{\mathbb{P}_\pi\left[x_s=j \text{ and } x_{s+1}=i\right]}{\mathbb{P}_\pi[x_{s+1}=i]} \\
&= \frac{\mathbb{P}_\pi[x_s=j]\mathbb{P}_j[x_1=i]}{\mathbb{P}_\pi[x_{s+1}=i]} \\
&= \frac{\pi_j a_{ji}}{\pi_i}.
\end{aligned}$$

Thus, we see that if the initial distribution is equal to the stationary distribution, the reverse is, indeed, the original Markov chain being run "backward." This can be shown to be the case for an arbitrary initial distribution, although the calculations are more involved.

We note that if $\Pi$ is the diagonal matrix with $i$th diagonal entry equal to $\pi(i)$, the reverse of $X$ is the Markov chain associated with the matrix

$$A^R = \Pi^{-1} A^T \Pi.$$

Thus, reversible Markov chains are simply those that are identical to their reverse – the matrix $\Pi A$ is symmetric if and only if $A = \Pi^{-1} A^T \Pi$.

It is clear that the stationary distribution of the reverse is identical to that of the original matrix: since $\Pi \mathbb{1} = \pi$,

$$\pi^T A^R = \left(\mathbb{1}^T \Pi\right)\left(\Pi^{-1} A^T \Pi\right) = \mathbb{1}^T A^T \Pi = \mathbb{1}^T \Pi = \pi^T.$$

Let the Markov chain $X$ be reducible. If $X$ has no transient states we may still

define the reverse $X^R$. Let $A$ be the transition matrix of $X$. Since $X$ has no transient states, $A$ is permutation similar to a block diagonal matrix:

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{bmatrix},$$

where each $A_k$ is an irreducible stochastic matrix. We form the reverse of $X$ by simply forming the reverses of its irreducible components. The reverse of $X$ is the Markov chain associated with the matrix

$$A^R \cong \begin{bmatrix} A_1^R & & 0 \\ & \ddots & \\ 0 & & A_m^R \end{bmatrix}.$$

### 2.4.2 Random walks

We present a graph theoretic definition of reversible Markov chains.

An undirected graph $G$ is an ordered pair $G = (V, E)$. The finite collection $V$ is the vertex set of $G$ and is generally assumed to be equal to $\{1, \ldots, n\}$ for some positive integer $n$. The edge set $E$ is a collection of unordered pairs of vertices. We express a given element of $E$ as $ij$ (where $i$ and $j$ are vertices); thus, $ij$ and $ji$ refer to the same edge in a given graph. A loop is an edge that has its endpoints identical; that is, an edge of the form $ii$. An *isolated vertex* in a graph is a vertex that is not the endpoint of any edge. (A vertex that has a loop but is incident to no other edges is

not considered isolated.) In this section we only consider graphs that have no isolated vertices. A *weighted graph* is a graph $G = (V, E)$ together with a function $w$ that maps the elements of $E$ to positive real numbers; we extend $w$ to a function on $V \times V$ by assigning the value $w_{ij} = 0$ for all $ij \notin E$.

**Definition 2.15.** Let $G$ be a weighted graph with vertices $V$ and weight $w$ that has no isolated vertices. The *random walk* on $G$ is the Markov chain with state space $\mathcal{S} = V$ and transition matrix $A$ given by

$$a_{ij} = \frac{w_{ij}}{\sum\limits_{k \in \mathcal{S}} w_{ik}}.$$

The matrix $A$, above, is clearly a stochastic matrix. Since there are no isolated vertices, the sum of the weights of the edges incident to a given vertex is not 0; and, for each $i$,

$$\sum_{j \in \mathcal{S}} a_{ij} = \sum_{j \in \mathcal{S}} \frac{w_{ij}}{\sum\limits_{k \in \mathcal{S}} w_{ik}} = \frac{\sum\limits_{j \in \mathcal{S}} w_{ij}}{\sum\limits_{k \in \mathcal{S}} w_{ik}} = 1.$$

**Theorem 2.16.** *Let $X$ be a time-homogeneous Markov chain on finite state space $\mathcal{S}$ with discrete time parameter. Then, $X$ is reversible if and only if it can be expressed as a random walk on a weighted graph with no isolated vertices.*

**Proof** Let $G$ be a weighted graph with no isolated vertices and let $X$ be the random walk on $G$. Let $W$ be the matrix corresponding to the weight function on $G$; the matrix $W$ is symmetric and nonnegative. Let $A$ be the transition matrix of $X$ and $\Pi$ be the diagonal matrix with

$$\pi_i = \sum_k w_{ik}.$$

Then, we simply note that $A = \Pi^{-1}W$, implying $W = \Pi A$. The stochastic matrix $A$ is reversible, as $W$ is symmetric and $\Pi$ has positive diagonal entries.

Now, let $X$ be a reversible Markov chain on state space $\mathcal{S}$ and let $A$ be the transition matrix of $X$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Let $w_{ij} = (\Pi A)_{ij} = \pi_i a_{ij}$; let $G$ be the weighted graph with $V = \mathcal{S}$,

$$E = \{ij : w_{ij} \neq 0\}$$

and weight $w$. Clearly, the random walk on $G$ is identical to $X$. ∎

Let $W$ be the adjacency matrix of a weighted graph (with no isolated vertices). By the above, the transition matrix of the associated random walk is $\Pi^{-1}W$, where $\Pi$ is the unique diagonal matrix such that $\Pi\mathbb{1} = W\mathbb{1}$. Further, the vector $\pi = \Pi\mathbb{1}$ is a scalar multiple of a stationary distribution of the associated random walk.

# Chapter 3

# Nearly uncoupled Markov chains

We present the concept of a nearly uncoupled Markov chain. Algorithmic analysis of such Markov chains will be the focus of the remainder of this work.

## 3.1 Definition

**Definition 3.1.** Let $A$ be an irreducible stochastic matrix on the state space $\mathcal{S}$, let $\pi$ be the stationary distribution of $A$ and let $\mathcal{E} \subseteq \mathcal{S}$. We define the $\pi$-coupling measure of $\mathcal{E}$ to be the value

$$w_\pi(\mathcal{E}) = \frac{\sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{E}} \pi_i a_{ij}}{\sum_{i\in\mathcal{E}} \pi_i}$$

and we define the $\mathbb{1}$-coupling measure of $\mathcal{E}$ to be

$$w_{\mathbb{1}}(\mathcal{E}) = \frac{\sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{E}} a_{ij}}{|\mathcal{E}|}.$$

As well, let $B = A(\mathcal{E})$; then, we define $w_\pi(B) = w_\pi(\mathcal{E})$ and $w_{\mathbb{1}}(B) = w_{\mathbb{1}}(\mathcal{E})$. If

the stochastic matrix $A$ is reducible, the $\pi$-coupling measure is undefined and the $\mathbb{1}$-coupling measure is as above.

Let $A$, $\mathcal{S}$ and $\pi$ be as in Definition 3.1. For a collection $\mathcal{E} \subseteq \mathcal{S}$, and a state $i \in \mathcal{E}$,

$$\sum_{j \in \mathcal{E}} a_{ij} = \sum_{j \in \mathcal{E}} \mathbb{P}\left[x_{t+1} = j | x_t = i\right] = \mathbb{P}\left[x_{t+1} \in \mathcal{E} | x_t = i\right].$$

So, for any $\mathcal{E} \subseteq \mathcal{S}$, the value $w_{\mathbb{1}}(\mathcal{E})$ is simply the average probability of transitioning from a state in $\mathcal{E}$ to another:

$$w_{\mathbb{1}}(\mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbb{P}\left[x_{t+1} \in \mathcal{E} | x_t = i\right].$$

The value $w_\pi(\mathcal{E})$ is a weighted average using weights determined by the vector $\pi$. By Proposition 1.5 and the fact that $\pi^T A = \pi^T$,

$$\pi_i = \mathbb{P}_\pi\left[x_t = i\right],$$

for all $t \geq 0$. Thus,

$$
\begin{aligned}
w_\pi(\mathcal{E}) &= \frac{\sum_{i\in\mathcal{E}} \mathbb{P}_\pi[x_t=i]\mathbb{P}[x_{t+1}\in\mathcal{E}|x_t=i]}{\sum_{i\in\mathcal{E}} \mathbb{P}_\pi[x_t=i]} \\
&= \frac{\sum_{i\in\mathcal{E}} \mathbb{P}_\pi[x_t=i]\mathbb{P}[x_{t+1}\in\mathcal{E}|x_t=i]}{\mathbb{P}_\pi[x_t\in\mathcal{E}]} \\
&= \sum_{i\in\mathcal{E}} \frac{\mathbb{P}_\pi[x_t=i]}{\mathbb{P}_\pi[x_t\in\mathcal{E}]} \mathbb{P}\left[x_{t+1}\in\mathcal{E}|x_t=i\right] \\
&= \sum_{i\in\mathcal{E}} \mathbb{P}_\pi\left[x_t=i|x_t\in\mathcal{E}\right] \mathbb{P}\left[x_{t+1}\in\mathcal{E}|x_t=i\right] \\
&= \mathbb{P}_\pi\left[x_{t+1}\in\mathcal{E}|x_t\in\mathcal{E}\right].
\end{aligned}
$$

If the initial distribution of the Markov chain associated with $A$ is distinct from its stationary distribution, we still have the weaker condition

$$\pi_i = \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{P}_v\left[x_s=i\right],$$

regardless of the initial distribution $v$ (this is a consequence of Theorem 2.2). So, the $\pi$-coupling measure $w_\pi(\mathcal{E})$ can be viewed as the long-term expected value of the probability of transitioning from $\mathcal{E}$ to $\mathcal{E}$. We interpret the $\mathbb{1}$-coupling measure as a short-term probability of transitioning from $\mathcal{E}$ to $\mathcal{E}$.

**Definition 3.2.** Let $B$ be a substochastic matrix. We define the *error vector* of $B$ to be the vector

$$\gamma_B = \mathbb{1} - B\mathbb{1} = (I - B)\,\mathbb{1}.$$

The value $\gamma_B(i)$ is referred to as the *error* at state $i$. The value

$$\eta(B) = \mathbb{1}^T \gamma_B = \mathbb{1}^T (I - B)\,\mathbb{1}$$

is the *total error* of $B$.

The error vector of a substochastic matrix $B$ is a measure of how close $B$ is to being stochastic – it measures how different each row sum is from 1.

Let $X$ be a Markov chain with state space $\mathcal{S}$ and transition matrix $A$; let $\mathcal{E} \subseteq \mathcal{S}$ be nonempty and let $B = A(\mathcal{E})$. Then, for $i \in \mathcal{E}$, we have

$$\gamma_B(i) = 1 - \sum_{j \in \mathcal{E}} a_{ij} = \sum_{j \notin \mathcal{E}} a_{ij} = \mathbb{P}\left[x_{t+1} \notin \mathcal{E} | x_t = i\right].$$

Thus, for each $i \in \mathcal{E}$, $\gamma_B(i)$ is the probability of transitioning from $i$ to a state not contained in $\mathcal{E}$.

**Definition 3.3.** Let $X$ be an discrete-time time-homogeneous Markov chain on a finite state space $\mathcal{S}$ with transition matrix $A$ and let $0 \leq \epsilon < 1$. Let $\mathcal{E} \subseteq \mathcal{S}$ be a nonempty proper subcollection of states and let $B = A(\mathcal{E})$. We refer to $\mathcal{E}$ as an *almost invariant aggregate with respect to* $\epsilon$, if $\gamma_B \leq \epsilon\mathbb{1}$.

Let $X$ be a Markov chain with state space $\mathcal{S}$. An almost invariant aggregate is a nonempty collection $\mathcal{E} \subseteq \mathcal{S}$ such that if $x_t \in \mathcal{E}$ then the probability that $x_{t+1} \notin \mathcal{E}$ is

less than or equal to $\epsilon$. If $\mathcal{E}$ is such an aggregate and $x_0 \in \mathcal{E}$, then the probability that $x_1, \ldots, x_t \in \mathcal{E}$ is greater than or equal to $(1-\epsilon)^t$. An almost invariant aggregate is a collection of states $\mathcal{E}$ such that when the Markov chain enters $\mathcal{E}$, it tends to remain in $\mathcal{E}$ for relatively long periods of time.

**Definition 3.4.** Let $X$ be a discrete time, time-homogeneous Markov chain on a finite state space $\mathcal{S}$ with transition matrix $A$ and let $0 \leq \epsilon < 1$. We refer to $X$ and $A$ as *nearly uncoupled with respect to* $\epsilon$ if $\mathcal{S}$ contains two or more disjoint almost invariant aggregates with respect to $\epsilon$.

We use the same definition of an almost invariant aggregate as [19]. In [7, 8], the collection $\mathcal{E} \subseteq \mathcal{S}$ is defined to be almost invariant if $w_\pi(\mathcal{E}) \geq 1 - \epsilon$. In [10], the authors propose that the condition $w_{\mathbb{1}}(\mathcal{E}) \geq 1 - \epsilon$ is a more useful criterion (than that in [7, 8]) for an almost invariant aggregate. The authors of [10] claim that the difficulties involved in solving the eigenvector equation $x^T A = x^T$ for $x$ make an approach that does not involve the stationary distribution more robust.

In Section 4.3, we present an example of how the property of being nearly uncoupled can effect the convergence of a Markov chain to its stationary distribution. In light of this, we propose that the $\pi$-coupling measure should only be used if the stationary distribution is known in advance. (In Appendix C we make use of the $\pi$-coupling measure in just such as case.)

When the value $\epsilon$ has been clearly specified, or its exact value is not of particular

interest, we will simply use the terms almost invariant aggregate and nearly uncoupled Markov chain.

We note that Definition 3.1 is the most conservative of the three. Let $A$ be a stochastic matrix and let $B = A(\mathcal{E})$ be a principal submatrix of $A$ such that $\gamma_B \leq \epsilon\mathbb{1}$. Then, for all $i \in \mathcal{E}$,

$$\mathbb{P}\left[x_{t+1} \in \mathcal{E} | x_t = i\right] = 1 - \gamma_B(i) \geq 1 - \epsilon.$$

So, any weighted average of these values, including $w_\pi(\mathcal{E})$ and $w_{\mathbb{1}}(\mathcal{E})$, is bounded below by $1 - \epsilon$.

**Lemma 3.5.** *Let $A$ be a stochastic matrix. Then, $A$ is nearly uncoupled with respect to $0 \leq \epsilon < 1$ if and only if $A$ can be expressed as a matrix of the form*

$$A \cong \begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mm} \end{bmatrix} \text{ or } \begin{bmatrix} A_{11} & \cdots & A_{1m} & A_{10} \\ \vdots & \ddots & \vdots & \vdots \\ A_{m1} & \cdots & A_{mm} & A_{m0} \\ A_{01} & \cdots & A_{0m} & A_{00} \end{bmatrix},$$

*where $m \geq 2$ and the following 3 conditions hold:*

1. *for $k = 1, \ldots, m$, $\gamma_{A_{kk}} \leq \epsilon\mathbb{1}$;*
2. *for $k = 1, \ldots, m$, the only principal submatrix of $A_{kk}$ satisfying condition 1 is $A_{kk}$ itself; and*

3. *when the matrix $A_{00}$ is present, no principal submatrix of $A_{00}$ (including $A_{00}$, itself) satisfies condition* 1.

**Proof** Clearly, a matrix of the above form is nearly uncoupled (the collections of states corresponding to the first $m$ blocks are disjoint almost invariant aggregates).

Now, suppose that $A$ is nearly uncoupled with respect to $\epsilon$. Then, we can express $A$ as an $m \times m$ or $(m+1) \times (m+1)$ block matrix, for some integer $m \geq 2$, such that the first $m$ principal submatrices on the diagonal satisfy condition 1 (we simply let the first $m$ blocks be principal submatrices of some collection of disjoint almost invariant aggregates). There is an upper limit as to how large $m$ can be in such an expression, namely, the order of $A$. So, suppose that $A$ has been expressed as an $m \times m$ or $(m+1) \times (m+1)$ block matrix satisfying condition 1 and that the integer $m$ is maximal among all such expressions of $A$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0$ be the collections of states corresponding to this expression. We note that we may have $\mathcal{E}_0 = \emptyset$.

For $k = 1, \ldots m$, the principal submatrix $A_{kk}$ may not satisfy condition 2. If $A_{kk}$ does satisfy condition 2, let $\mathcal{E}'_k = \mathcal{E}_k$; otherwise, let $A'_{kk}$ be a principal submatrix of $A_{kk}$ such that

1. $\gamma_{A'_{kk}} \leq \epsilon \mathbb{1}$, and

2. among all principal submatrices $B$ of $A_{kk}$ such that $\gamma_B \leq \epsilon \mathbb{1}$, $A'_{kk}$ has minimal order.

Such a submatrix must exist, as $A_{kk}$ itself has $\gamma_{A_{kk}} \leq \epsilon\mathbb{1}$. Thus, the only principal submatrix of $A'_{kk}$ satisfying condition 1 is $A'_{kk}$. Let $\mathcal{E}'_k$ be the collection of states corresponding to $A'_{kk}$.

Now, let

$$\mathcal{E}'_0 = \mathcal{S} \setminus \bigcup_{k=1}^{m} \mathcal{E}'_k$$

and, if $\mathcal{E}'_0 \neq \emptyset$, let $A'_{00} = A(\mathcal{E}'_0)$. If we now suppose that there is a principal submatrix $B$ of $A'_{00}$ such that $\gamma_B \leq \epsilon\mathbb{1}$, this contradicts the maximality of $m$ (we could then construct an $(m+1)$th almost invariant aggregate). Thus, the collections $\mathcal{E}'_1, \ldots, \mathcal{E}'_m$ are a set of minimal disjoint almost invariant aggregates and the collection $\mathcal{E}'_0$ (when it is nonempty) does not contain an almost invariant aggregate. The expression of $A$ with $A_{ij} = A(\mathcal{E}'_i, \mathcal{E}'_j)$ satisfies the statement of the theorem. ∎

The states corresponding to block index 0 in Lemma 3.5 represent states that are not part of any almost invariant aggregate. They are states that the Markov chain visits only rarely, as the probability of entering this collection is at most $\epsilon$, but the probability of leaving it is strictly higher. We will refer to the members of this collection, when it is nonempty, as *near transient* states.

We will use Lemma 3.5 as the canonical representation of the transition matrix of a nearly uncoupled Markov chain. Let $X$ be a nearly uncoupled Markov chain, with respect to $\epsilon$, on the state space $\mathcal{S}$. An *$\epsilon$-uncoupling* of $X$ is a partition

$$\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$$

of $\mathcal{S}$, where $m \geq 2$ and $\mathcal{E}_0$ is allowed to be empty, such that, for $k \neq 0$, $\mathcal{E}_k$ is a minimal almost invariant aggregate and, if it is nonempty, $\mathcal{E}_0$ does not contain any almost invariant aggregates as subsets. (We specify that $\mathcal{E}_0$ may be empty as, normally, the members of a partition are nonempty.) We note that if $A$ is the transition matrix of $X$ and $\mathcal{E}_k$ is a minimal almost invariant aggregate, the principal matrix $A(\mathcal{E}_k)$ must be irreducible.

**Example 3.6.** The matrix $A_1$ is nearly uncoupled with respect to $\epsilon = 0.05$. Its minimal almost invariant aggregates are $\mathcal{E}_1 = \{1, 2\}$ and $\mathcal{E}_2 = \{3\}$. It possesses one near transient state; namely state 4.

$$A_1 = \begin{bmatrix} \mathbf{0.34} & \mathbf{0.62} & 0.03 & 0.01 \\ \mathbf{0.21} & \mathbf{0.77} & 0.02 & 0 \\ 0.01 & 0.03 & \mathbf{0.95} & 0.01 \\ 0.21 & 0.19 & 0.32 & 0.28 \end{bmatrix}.$$

The matrix $A_2$ is nearly uncoupled with respect to $\epsilon = 0.05$. We note that its decomposition into almost invariant aggregates and near transient states is not unique.

$$A_2 = \begin{bmatrix} 0.94 & 0.02 & 0.03 & 0.01 \\ 0.03 & 0.93 & 0.03 & 0.01 \\ 0.03 & 0.04 & 0.92 & 0.01 \\ 0.01 & 0.01 & 0.02 & 0.96 \end{bmatrix}.$$

The collections $\{1,2\}$, $\{1,3\}$, $\{2,3\}$ and $\{4\}$ are all minimal almost invariant aggregates, with respect to $\epsilon = 0.05$. There are three different possibilities for the representation of this matrix described in Lemma 3.5. In this particular example, we can see that we simply chose $\epsilon$ poorly – a unique decomposition arises for $\epsilon = 0.08$, for example. However, without knowing the structure of the aggregates *a priori,* one may not be able to choose $\epsilon$ to produce such a nice structure. As well, the particular application may force the choice of $\epsilon$. So, we cannot typically assume that the matrix has a unique similarity to a near block lower-triangular form.

In [13], a much more strict definition of a nearly uncoupled stochastic matrix is presented. Let $A$ be a stochastic matrix and let $\delta \geq 0$. The authors consider a matrix $A$ to be nearly uncoupled with respect to $\delta$ if there is a partition $\Psi = \{\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0\}$, where $m \geq 2$ and $\mathcal{E}_0$ is allowed to be empty, such that the *coupling measure* $\mu$, defined by

$$\mu(\Psi) = \sum_{k=1}^{m} \sum_{i \in \mathcal{E}_k} \sum_{j \notin \mathcal{E}_k} a_{ij},$$

has $\mu(\Psi) \leq \delta$. In terms of error vectors and total error, as we have defined them,

$$\mu(\Psi) = \sum_{k=1}^{m} \mathbb{1}^T \gamma_{A(\mathcal{E}_k)} = \sum_{k=1}^{m} \eta(A(\mathcal{E}_k)).$$

Let $A$ be nearly uncoupled stochastic matrix of order $n$ with respect to $\epsilon$ (under Definition 3.4) and let $\Psi$ be an uncoupling as in Lemma 3.5. Then,

$$\mu(\Psi) \leq \epsilon \left(n - |\mathcal{E}_0|\right) \leq \epsilon n.$$

So, if a Markov chain with a state space of $n$ states is nearly uncoupled with respect to $\epsilon$, via Definition 3.4, then it is nearly uncoupled with respect to $\delta = \epsilon n$, via the $\mu$-criterion.

## 3.2 Problem statement

We are interested in solving the following problem. Let $\tilde{A}$ be a stochastic matrix of the form given in Lemma 3.5 with $m \geq 2$, let $P$ be an arbitrary permutation matrix of the same order as $\tilde{A}$ and let $A = P\tilde{A}P^T$. Without *a priori* knowledge of the matrix $P$, can we recover the uncoupled structure of $\tilde{A}$? That is, if a given matrix is nearly uncoupled, but states from distinct almost invariant aggregates have been "scrambled" together, can we reorder the states so that the near block diagonal structure is apparent? Moreover, can we produce such a reordering without knowing, in advance, whether or not $A$ even has such a structure?

In attempting to solve this problem, it suffices to produce an $\epsilon$-uncoupling

$$\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$$

of the state space $\mathcal{S}$ of $A$.

Suppose that $A$ is nearly uncoupled and that $\Psi$ is such a partition of the associated state space. Let $n_k = |\mathcal{E}_k|$. We list the elements of $\mathcal{S}$ as a sequence $i_1, i_2, \ldots, i_n$ such that

$$\mathcal{E}_1 = \{i_1, \ldots, i_{n_1}\}, \ \mathcal{E}_2 = \{i_{n_1+1}, \ldots, i_{n_1+n_2}\},$$

and so forth. Then, we let $P$ be the permutation matrix such that for each $k$, the $i_k$th entry in the $k$th row is the unique entry equal to 1 in that row. *i.e.* For all $i$ and $j$,

$$p_{ij} = \begin{cases} 1 & \text{if } i = i_j, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$\mathcal{E}'_1 = \{1, \ldots, n_1\}, \ \mathcal{E}'_2 = \{n_1 + 1, \ldots, n_1 + n_2\},$$

and so forth. The matrix $\tilde{A} = PAP^T$ then has

$$\tilde{A}(\mathcal{E}'_k) \cong A(\mathcal{E}_k),$$

for each $k$.

Our solutions (the stochastic complement based algorithms) will be purely constructive in nature. Given a matrix $A$, we attempt to produce a collection of disjoint almost invariant aggregates. If two or more almost invariant aggregates are constructed, then we have achieved our goal; if only one aggregate is produced, either the matrix is not nearly uncoupled or our method has failed.

We are particularly interested in methods which avoid using the eigenvalue and/or singular value decomposition of a matrix altogether. This is for two reasons. Firstly, there is already a wealth of research concerned with solving this problem using spectral or singular value based methods. Secondly, we suspect that in extreme cases, the eigenvectors associated with eigenvalues near to 1 may be difficult to calculate and possibly overly sensitive to perturbations.

## 3.3 Perron cluster based algorithms

In [7, 8], two algorithms which attempt to produce almost invariant aggregates of a reversible Markov chain are presented. We present a brief description of the algorithm referred to as Perron cluster cluster analysis (PCCA); see [7] for a thorough exposition.

Let $\mathcal{S}$ be a collection of indices/states and let $\mathcal{C} \subseteq \mathcal{S}$. The *characteristic vector* of $\mathcal{C}$, labelled $\mathbb{1}_{\mathcal{C}}$, is the $(0,1)$-vector on $\mathcal{S}$ that has its $i$th entry equal to 1 if $i \in \mathcal{C}$ or 0 if $i \notin \mathcal{C}$. A *characteristic collection* is a set of characteristic vectors (none of which

is the 0-vector) whose sum is $\mathbb{1}$. A characteristic collection corresponds to a unique partition of $\mathcal{S}$. Moreover, a characteristic collection is clearly linearly independent.

The authors only consider decompositions which do not include near transient states. That is, if the matrix $A$ is nearly uncoupled with respect to $\epsilon$, the assumption is made that there is a partition $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m)$ of the associated state space that has $\gamma_{A_l} \leq \epsilon\mathbb{1}$, for each of the principal submatrices $A_l = A(\mathcal{E}_l)$. Let $A$ be a nearly uncoupled reversible stochastic matrix and let $\Psi$ be such an uncoupling. Without loss of generality, we assume that

$$A = \begin{bmatrix} A_{11} & & A_{ij} \\ & \ddots & \\ & & A_{mm} \end{bmatrix},$$

where each $B = A_{ll}$ has $\gamma_B \leq \epsilon\mathbb{1}$. Let

$$\Pi = \begin{bmatrix} \Pi_1 & & 0 \\ & \ddots & \\ 0 & & \Pi_m \end{bmatrix}$$

be a positive diagonal matrix such that

$$W = \Pi A = \begin{bmatrix} \Pi_1 A_{11} & & \Pi_i A_{ij} \\ & \ddots & \\ & & \Pi_m A_{mm} \end{bmatrix} = \begin{bmatrix} W_{11} & & W_{ij} \\ & \ddots & \\ & & W_{mm} \end{bmatrix}$$

is symmetric. We further assume that $\mathbb{1}^T \Pi \mathbb{1} = 1$, so that $\Pi\mathbb{1} = \pi$ is the stationary distribution of $A$. Since $\Pi$ is a positive diagonal matrix, for any real number $z$, the

matrix exponent $\Pi^z$, defined to be the positive diagonal matrix with its $i$th diagonal entry equal to $\pi_i^z$, exists. We note that the matrix $\Pi^{1/2}A\Pi^{-1/2} = \Pi^{-1/2}W\Pi^{-1/2}$ is symmetric and so $A$ has real eigenvalues. Let $A_D$ be the block-diagonal matrix with the same diagonal blocks $A_D(\mathcal{E}_l) = A(\mathcal{E}_l)$ as $A$ and let $A_O = A - A_D$:

$$A_D = \begin{bmatrix} A_{11} & & 0 \\ & \ddots & \\ 0 & & A_{mm} \end{bmatrix} \text{ and } A_O = \begin{bmatrix} 0 & & A_{ij} \\ & \ddots & \\ & & 0 \end{bmatrix}.$$

That is, $A_)(\mathcal{E}_k, \mathcal{E}_l) = A(\mathcal{E}_k, \mathcal{E}_l)$ if $k \neq l$ and $A_O(\mathcal{E}_k) = 0$ for all $k$.

Since $\gamma_{A_{ll}} = (I - A_{ll})\mathbb{1} \leq \epsilon\mathbb{1}$, the Perron value of each $A_{ll}$ is greater than or equal to $1-\epsilon$; this implies that $A_D$ has at least $m$ eigenvalues greater than or equal to $1-\epsilon$. As well, $A_0\mathbb{1} \leq \epsilon\mathbb{1}$ and the matrix $A_0$ is entrywise nonnegative, so any eigenvalue of $A_0$ is contained in the interval $[-\epsilon, \epsilon]$ (via the Perron-Frobenius theorem).

According to [14, Theorem 4.3.1], if $B$ and $C$ are symmetric matrices with eigenvalues

$$\lambda_1(B) \geq \ldots \geq \lambda_n(B) \text{ and } \lambda_1(C) \geq \ldots \geq \lambda_n(C),$$

the eigenvalues of $B + C$,

$$\lambda_1(B + C) \geq \ldots \geq \lambda_n(B + C),$$

satisfy

$$\lambda_k(B + C) \geq \lambda_k(B) + \lambda_n(C),$$

for $1 \leq k \leq n$. The $m$ largest eigenvalues of $A_D$ are greater than or equal to $1-\epsilon$ and the smallest eigenvalue of $A_0$ is greater than or equal to $-\epsilon$. The matrices $A_D$ and $A_O$ are not necessarily symmetric, but the matrices $\Pi^{1/2}A_D\Pi^{-1/2}$ and $\Pi^{1/2}A_O\Pi^{-1/2}$ are. So, the $m$ largest eigenvalues of

$$\Pi^{1/2}A_D\Pi^{-1/2} + \Pi^{1/2}A_O\Pi^{-1/2} = \Pi^{1/2}A\Pi^{-1/2}$$

are each greater than or equal to $1-2\epsilon$. Thus, $A$ has $m$ eigenvalues that are greater than or equal to $1-2\epsilon$. We refer to the eigenvalues nearest to 1 of a reversible nearly uncoupled stochastic matrix as the *Perron cluster*.

For each $l$, let $D_l$ be the nonnegative diagonal matrix satisfying $D_l\mathbb{1} = (I - A_{ll})\mathbb{1}$ and let $\tilde{A}_{ll} = A_{ll} + D_l$. Let

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & & 0 \\ & \ddots & \\ 0 & & \tilde{A}_{mm} \end{bmatrix}.$$

We note that if $i \in \mathcal{E}_l$, then

$$|\tilde{a}_{ij} - a_{ij}| = \begin{cases} 0 & \text{if } j \in \mathcal{E}_l \text{ and } j \neq i \\ \sum_{k \notin \mathcal{E}_l} a_{ik} & \text{if } j = i \\ a_{ij} & \text{if } j \notin \mathcal{E}_l. \end{cases}$$

So,

$$\begin{aligned}
\left\|\Pi(\tilde{A}-A)\right\|_\infty &= \max_{i\in\mathcal{S}}\left\{\sum_{j\in\mathcal{S}}\left|\pi_i\left(\tilde{a}_{ij}-a_{ij}\right)\right|\right\} \\
&= \max_{1\le l\le m}\max_{i\in\mathcal{E}_l}\left\{\sum_{j\in\mathcal{S}}\left|\pi_i\left(\tilde{a}_{ij}-a_{ij}\right)\right|\right\} \\
&= \max_{1\le l\le m}\max_{i\in\mathcal{E}_l}\left\{2\sum_{j\notin\mathcal{E}_l}\pi_i a_{ij}\right\} \\
&\le \max_{1\le l\le m}\max_{i\in\mathcal{E}_l}\left\{2\frac{\sum_{j\notin\mathcal{E}_l}\pi_i a_{ij}}{\sum_{j\in\mathcal{E}_l}\pi_j}\right\} \\
&= \max_{1\le l\le m}\left\{2\left(1-w_\pi(\mathcal{E}_l)\right)\right\} \\
&\le 2\epsilon.
\end{aligned}$$

The matrix $\tilde{A}$ has at least $m$ eigenvalues equal to 1. We further assume that $\tilde{A}$ has exactly $m$ eigenvalues equal to 1 and that the remainder are bounded (in absolute value) by some number significantly less than 1 [7, 8, 10, 18]. In the next section, we examine some of the implications of this assumption.

By Theorem 2.5, the (right) eigenspace of $\tilde{A}$ associated with eigenvalue 1 is the linear span of the collection $\{\mathbb{1}_{\mathcal{E}_l}\}$.

Now, suppose that $Av = \lambda v$ where $\lambda \ge 1-2\epsilon$ and $\|v\|_\infty = 1$. Then, since $(A-\lambda I)v = 0$,

$$\begin{aligned}
\|\Pi(\tilde{A}-I)v\|_\infty &\le \|\Pi(A-\lambda I)v\|_\infty + \|\Pi(\tilde{A}-A)v\|_\infty + (1-\lambda)\|\Pi v\|_\infty \\
&\le 0 + \|\Pi(\tilde{A}-A)\|_\infty\|v\|_\infty + (1-\lambda)\|\Pi\|_\infty\|v\|_\infty \\
&\le 2\epsilon + (1-\lambda) \\
&\le 4\epsilon.
\end{aligned}$$

We conclude that $v$ must be a small perturbation of some member of the span of $\{\mathbb{1}_{\mathcal{E}_l}\}$ (as $\tilde{A}$ has no other eigenspaces associated with eigenvalues near 1). We note that this implies that $|v_i - v_j|$ is small whenever $i$ and $j$ are contained in the same aggregate $\mathcal{E}_l$.

So, suppose that $A$ and the other relevant terms are as above. Let $v_1, \ldots, v_m$ be a collection of eigenvectors associated with the Perron cluster. If $|v_k(i) - v_k(j)|$ is small for all $k$ we conclude that the states $i$ and $j$ are contained in the same almost invariant aggregate. If there is at least one $l$ with $|v_l(i) - v_l(j)|$ large (relatively) we conclude that $i$ and $j$ are members of distinct almost invariant aggregates.

Utilizing these ideas, we have a sketch of the PCCA algorithm; the algorithm takes as inputs a nearly uncoupled reversible stochastic matrix $A$ of order $n$ (which is assumed to have no near transient states and fast-mixing almost invariant aggregates), and the Perron cluster $\{\lambda_1, \ldots, \lambda_m\}$ of $A$. The Perron cluster can be identified by examining the eigenvalues of $A$ to see if there is a clear cluster about 1, or by simply choosing an arbitrary small value $\delta > 0$ and then selecting those eigenvalues with $|1 - \lambda| < \delta$.

**Algorithm 1** The Perron cluster cluster analysis algorithm

1. Let $\{v_1, \ldots, v_m\}$ be right eigenvectors of $A$ associated with the Perron cluster.
2. Produce a characteristic collection of vectors $\{w_1, \ldots, w_m\}$ such that if $w_k(i) = w_k(j) = 1$ for some $k$, then $|v_l(i) - v_l(j)|$ is small for all $l$.
3. Return the aggregates $\mathcal{E}_k = \{i : w_k(i) = 1\}$.

Step 2 is a somewhat involved process; see [7, 8] for the details.

A *near-characteristic collection* is a set of column vectors $\{w_k\}$ on the index set $\mathcal{S}$ which are entrywise nonnegative and whose sum is $\mathbb{1}$. *i.e.* Each $w_k \geq 0$ and for all $i \in \mathcal{S}$,

$$\sum_{k=1}^{m} w_k(i) = 1.$$

A near-characteristic collection can be used to partition its associated indices in much the same manner as a characteristic collection. We let

$$\mathcal{E}_k = \{i : \text{for all } l \neq k, w_k(i) \geq w_l(i)\}.$$

In other words, $\mathcal{E}_k$ is the collection of indices that are given the largest weight by $w_k$. If we do not see repeated values in the entries of the near-characteristic collection, the collections $\mathcal{E}_k$ form a partition; however, if state $i$ attains its maximum $w_k(i)$ for multiple values of $k$, we can simply arbitrarily assign it to one such $\mathcal{E}_k$ or use some other metric to decide on its place in a partition.

Near characteristic collections allow for a more detailed analysis of almost invariant aggregates. Let $\{w_k\}$ and $\{\mathcal{E}_k\}$ be as above. If $w_k(i)$ is large and for all $l \neq k$ the value $w_l(i)$ is insignificant, we view $i$ as being part of the "centre" of $\mathcal{E}_k$. If $w_k(i)$ and $w_l(i)$ are both significant, we view $i$ as being near the "border" between $\mathcal{E}_k$ and $\mathcal{E}_l$.

In this manner, a near-characteristic collection can be seen as a "fuzzy" partition, of sorts. Given a Markov chain $X$ on state space $\mathcal{S}$ and a near characteristic collection

of vectors $w_1, \ldots, w_m$, we define fuzzy collections $\mathcal{C}_1, \ldots, \mathcal{C}_m$ and say that if $x_t = i$, then the probability that $x_t \in \mathcal{C}_k$ is $w_k(i)$.

In [8], the algorithm PCCA+ is presented. It proceeds in the same manner as PCCA, except in step 2 a near-characteristic collection $\{w_k\}$ is produced. The PCCA+ algorithm attempts to construct $\{w_k\}$ in such a way that whenever $|w_k(i) - w_k(j)|$ is small for all $k$, $|v_k(i) - v_k(j)|$ is also small for all $k$. The authors state that algorithm PCCA+ is suspected to be the more robust of the two.

## 3.4 Fiedler vectors and connectivity

The Perron cluster approach is very much related to a concept known as the Fiedler vector.

Let $G$ be a connected weighted graph (with no isolated vertices) on the vertex set $\mathcal{S}$ and let $W$ be the symmetric matrix of weights associated with $G$. Let $D$ be the positive diagonal matrix with $i$th diagonal entry equal to

$$d_i = \sum_{j \in \mathcal{S}} w_{ij}.$$

Let $X$ be the random walk on $G$. So, the transition matrix of $X$ is the reversible stochastic matrix $A = D^{-1}W$ and the stationary distribution is the vector

$$\pi = \frac{1}{\sum\limits_{i \in \mathcal{S}} d_i} D\mathbb{1} = \frac{1}{\mathbb{1}^T D\mathbb{1}} D\mathbb{1}.$$

Since $D$ is a positive diagonal matrix, for any real number $z$, the matrix exponent $D^z$ exists. The *normalised Laplacian matrix* associated with $G$ is the symmetric matrix

$$L = D^{-1/2}\,(D - W)\,D^{-1/2} = I - D^{-1/2}WD^{-1/2}.$$

The normalised Laplacian is, necessarily, positive semidefinite (it is symmetric and each of its eigenvalues is nonnegative). Let

$$\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$$

be the eigenvalues of $L$. We note that $\lambda_0 = 0$ and that a null vector of $L$ is the vector $v = D^{1/2}\mathbb{1}$:

$$\left(D^{-1/2}\,(D - W)\,D^{-1/2}\right)\left(D^{1/2}\mathbb{1}\right) = D^{-1/2}\,(D - W)\,\mathbb{1} = 0.$$

(since $D\mathbb{1} = W\mathbb{1}$). The normalised Laplacian of $G$ is intimately related to the random walk on $G$, as the following matrix-similarity shows.

$$\begin{aligned} D^{1/2}AD^{-1/2} &= D^{1/2}\,(D^{-1}W)\,D^{-1/2} \\ &= D^{-1/2}WD^{-1/2} \\ &= I - L. \end{aligned}$$

So, given the eigenvalues $\{\lambda_k\}$ of the normalised Laplacian $L$ of $G$, the eigenvalues of the transition matrix $A$ of the random walk on $G$ are $\{1 - \lambda_k\}$.

Let $G$ be a weighted graph, let $L$ be the normalised Laplacian matrix of $G$ and let

$$0 = \lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$$

be the eigenvalues of $L$ (eigenvalues with multiplicities greater than 1 are included multiple times in this list). We will refer to an eigenvector associated with $\lambda_1$ as a *Fiedler vector* of $G$.

The Laplacian matrix of $G$ is defined by the formula $L' = D - W$, where $D$ and $W$ are as above. The Laplacian matrix and the normalised Laplacian matrix are connected via the formula

$$L' = D - W = D^{1/2}(I - D^{-1/2}WD^{-1/2})D^{1/2} = D^{1/2}LD^{1/2}.$$

See [4, Section 1.2] for a brief discussion concerning the differences between the Laplacian and the normalised Laplacian. We use the label Fiedler vector to refer to eigenvectors of the normalised Laplacian for simplicity's sake – generally, this label is only used to refer to eigenvectors of the Laplacian matrix.

The concept of a Fiedler vector first appeared in [9] (although, it is not there referred to as such). A Fiedler vector is related, in a very interesting manner, to connectivity properties of the graph $G$.

For example, suppose that $G$ is a tree and let $v$ be a Fiedler vector of the normalised Laplacian of $G$. Then, the induced subgraphs $G_+$ and $G_-$ on the vertex sets

$$\mathcal{S}_+ = \{i : v_i \geq 0\} \text{ and } \mathcal{S}_- = \{i : v_i < 0\}$$

are disjoint connected subtrees of $G$ joined by a single edge ([9, Theorem 3.14]).

The following proposition is a rewording of [4, Lemma 2.1]. (We have rephrased it in terms of transition probabilities; there it is presented in terms of edge weights.)

**Proposition 3.7.** *Let $A$ be an irreducible reversible stochastic matrix with stationary distribution $\pi$ on the state space $\mathcal{S}$; let $\mu$ be the eigenvalue of $A$ closest, but not equal, to 1. Let $\mathcal{E} \subseteq \mathcal{S}$ and let*

$$\alpha = \sum_{i \in \mathcal{E}} \pi_i.$$

*Then,*

$$w_\pi(\mathcal{E}) \leq 1 - \frac{1}{2}(1-\mu)(1-\alpha).$$

*Moreover, equality is attained if and only if there are positive constants $\beta$ and $\delta$ such that the vector $v$ with*

$$v_i = \begin{cases} \frac{\beta}{\sqrt{\pi_i}} & \text{if } i \in \mathcal{E} \\ -\frac{\delta}{\sqrt{\pi_i}} & \text{if } i \notin \mathcal{E} \end{cases}$$

*is a Fiedler vector of the associated normalised Laplacian.*

This is useful in that it provides a necessary, although not sufficient, condition for a matrix to be nearly uncoupled. Let $A$ be a reversible stochastic matrix on the

state space $\mathcal{S}$ and suppose that we are interested in determining whether or not the matrix is nearly uncoupled with respect to $\epsilon$. Assume that $A$ is nearly uncoupled (with respect to $\epsilon$), let $\pi$ be the associated stationary distribution and let $\mathcal{E}_1$ and $\mathcal{E}_2$ be disjoint almost invariant aggregates contained in the associated state space. It must be that for at least one of $l = 1$ or 2

$$\sum_{i \in \mathcal{E}_l} \pi_i \leq \frac{1}{2}.$$

Suppose that

$$\alpha = \sum_{i \in \mathcal{E}_1} \pi_i \leq \frac{1}{2}.$$

Then, we must have

$$1 - \epsilon \leq w_\pi(\mathcal{E}_1) \leq 1 - \frac{1}{2}(1 - \mu)(1 - \alpha) \leq 1 - \frac{1}{4}(1 - \mu),$$

further implying that $1 - 4\epsilon \leq \mu$.

So, if the eigenvalue $\mu$ of $A$ that is closest, but not equal, to 1 satisfies $\mu < 1 - 4\epsilon$, then we can conclude that $A$ is *not* nearly uncoupled, with respect to $\epsilon$. If, instead, we have $1 - 4\epsilon \leq \mu$, then it is entirely possible that $A$ is nearly uncoupled, with respect to $\epsilon$. Moreover, in this case, the Fiedler vector associated with normalised Laplacian gives us a potential starting point in attempting to construct almost invariant aggregates of $A$; one can begin by looking at the partition induced by the signs of the entries in the Fiedler vector.

The PCCA approach, in a sense, generalises this idea by considering what might be called "Fiedler spaces".

## 3.5 A singular value decomposition based algorithm

In [10], an alternate approach to constructing almost invariant aggregates of a given stochastic matrix is presented. Here, we only present the algorithm, itself. In [24], we examine this algorithm in detail and present some supplemental results concerning its implementation.

A *unitary* matrix is a square complex matrix $U$ such that $UU^* = I$. A real unitary matrix evidently satisfies $UU^T = I$; a real unitary matrix is referred to as an *orthogonal* matrix. Let $A$ be a $n \times n$ complex matrix. A singular value decomposition of $A$ is an expression

$$A = U\Sigma V^*$$

where $U$ and $V$ are unitary matrices and $\Sigma$ is a nonnegative diagonal matrix where the diagonal entries satisfy $\sigma_{ii} \geq \sigma_{jj}$ for all $i < j$. When $A$ is real then $U$ and $V$ can be taken to be orthogonal matrices, in which case we have

$$A = U\Sigma V^T.$$

The $i$th columns of $U$ and $V$ are referred to as left and right singular vectors, respectively, of $A$ associated with the singular value $\sigma_{ii}$. If $A$ is real and we let the $i$th

columns of $U$ and $V$ be $u_i$ and $v_i$, respectively, we then have

$$Av_i = \sigma_{ii}u_i \text{ and } A^T u_i = \sigma_{ii}v_i.$$

We label the singular values of $A$ as $\sigma_i(A) = \sigma_{ii}$. The number $\sigma_1(A)$ is, in fact, equal to the euclidean 2-norm of $A$: $\sigma_1(A) = \|A\|_2$. See [14] for a thorough exposition of the singular value decomposition.

The SVD-based algorithm is very simply expressed as a recursive algorithm. Its only input is a substochastic matrix $A$ on a state space $\mathcal{S}$. Within the algorithm there are references to singular vectors and coupling measures. It is up to the user to decide whether to utilise left or right-singular vectors and the $\pi$ or $\mathbb{1}$-coupling measure; we will use $w(\mathcal{E})$ to represent this undetermined coupling measure of the set $\mathcal{E}$. The output of the SVD-based algorithm is a partition $\{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$ of $\mathcal{S}$ such that $w(\mathcal{E}_k) > 1/2$ for all $k$.

**Algorithm 2** SVDA$(A, \mathcal{S})$

**if** $|\mathcal{S}| = 1$ **then**
  **return** $\{\mathcal{S}\}$
  Terminate the algorithm.
**end if**
Let $v$ be a singular vector associated with $\sigma_2(A)$.
Let $\mathcal{S}_+ = \{i \in \mathcal{S} : v_i \geq 0\}$, $\mathcal{S}_- = \{i \in \mathcal{S} : v_i < 0\}$, $A_+ = A(\mathcal{S}_+)$ and $A_- = A(\mathcal{S}_-)$.
**if** $w(A_+) \leq 1/2$ or $w(A_+) \leq 1/2$ **then**
  **return** $\{\mathcal{S}\}$
**else**
  **return** SVDA$(A_+, \mathcal{S}_+) \cup$ SVDA$(A_-, \mathcal{S}_-)$
**end if**

Algorithm 2 uses reasoning very similar to that of Proposition 3.7 – it simply uses

singular vectors rather than Fiedler vectors. An advantage of this approach is that singular vectors are, in general, more easily and reliably calculated than eigenvectors.

In [15], a somewhat similar algorithm is presented. Rather than examining a singular vector associated with the second largest singular value of $A$, this algorithm proceeds by examining a singular vector associated with the second smallest singular value of $I - A$ (where $A$ is the matrix in question).

# Chapter 4

# The stochastic complement

We present the stochastic complement, which will be our primary tool for constructing almost invariant aggregates of a given Markov chain. The stochastic complement is introduced in [19]. It is there utilised as a tool for constructing the stationary distribution of a Markov chain and analysing the rate of convergence of a Markov chain to its stationary distribution (see Section 4.3). Many of the results of this chapter are discussed in [19], although some appear without proof.

## 4.1 Definition

**Definition 4.1.** Let $A$ be a stochastic matrix with associated state space $\mathcal{S}$. Let $\{\mathcal{C}_1, \mathcal{C}_2\}$ be a partition of $\mathcal{S}$ and express

$$A \cong \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{ij} = A(\mathcal{C}_i, \mathcal{C}_j)$. If the matrix $I - A_{22}$ is nonsingular, we define the *stochastic complement* of $\mathcal{C}_1$ to be the matrix

$$S(\mathcal{C}_1) = A_{11} + A_{12}\left(I - A_{22}\right)^{-1} A_{21}.$$

If $I - A_{22}$ is singular, the stochastic complement of $\mathcal{C}_1$ is not defined.

**Proposition 4.2.** *Let $A$ be a stochastic matrix with state space $\mathcal{S}$ and let $\{\mathcal{C}_1, \mathcal{C}_2\}$ be a partition of $\mathcal{S}$. Then, the stochastic complement $S(\mathcal{C}_1)$ exists if and only if $\mathcal{C}_2$ does not contain an essential class of states.*

**Proof** Let $B = A(\mathcal{C}_2)$. The matrix $I - B$ is nonsingular if and only if 1 is not an eigenvalue of $B$. By Corollary 2.7, we see that $I - B$ is nonsingular if and only if $\mathcal{C}_2$ does not contain an essential class of states. ∎

**Remark.** The stochastic complement can be seen as a way of removing the states $\mathcal{C}_2$ from the associated Markov chain. Proposition 4.2 tells us that we cannot remove an entire essential class; *i.e.* the stochastic complement $S(\mathcal{C}_1)$ exists if and only if $\mathcal{C}_1$ contains at least one member of every essential class.

Let $A$ be a stochastic matrix and let

$$\omega = i_0 \to i_2 \to \cdots \to i_t$$

be a directed walk of length $t$ in the associated digraph. If $t \geq 1$, the *weight* of $\omega$ is the product of the $t$ transition probabilities along $\omega$:

$$a(\omega) = a_{i_0 i_1} a_{i_1 i_2} \cdots a_{i_{t-1} i_t}.$$

If $\omega$ is a walk of length 0, we define $a(\omega) = 1$. The weight $a(\omega)$ is the probability of transitioning from $i_0$ to $i_t$ via the walk $\omega$. By Proposition 1.5,

$$\left(A^t\right)_{ij} = \sum_{\omega \in \Omega_{ij}(t)} a(\omega),$$

where $\Omega_{ij}(t)$ is the collection of directed walks from $i$ to $j$ with length equal to $t$. Let $\omega$ be as above and let $\mathcal{C} \subset \mathcal{S}$ be a subcollection of the state space; if

$$i_1, \ldots, i_{t-1} \in \mathcal{C},$$

we refer to $\omega$ as a directed walk through $\mathcal{C}$,

$$\omega : i_0 \rightsquigarrow_{\mathcal{C}} i_t.$$

Note that the endpoints of a directed walk through $\mathcal{C}$ are not necessarily contained in $\mathcal{C}$; such a walk is merely one in which every interior point is contained in $\mathcal{C}$. Any directed walk of length 0 or 1 is trivially a walk through any collection, as it contains no interior points.

**Proposition 4.3.** *Let $A$ be a stochastic matrix and let $\mathcal{C}_1 \subseteq \mathcal{S}$ be a subcollection of the state space; let $\mathcal{C}_2 = \mathcal{S} \setminus \mathcal{C}_1$. If the stochastic complement $S(\mathcal{C}_1)$ is defined, it is itself a stochastic matrix and models the following Markov chain:*

1. *the state space is equal to* $\mathcal{C}_1$*; and*

2. *for* $i, j \in \mathcal{C}_1$*, the transition probability* $s_{ij}$ *is the sum of the weights of the directed walks*

$$i \rightsquigarrow_{\mathcal{C}_2} j.$$

**Proof** Let

$$A \cong \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

(as in 4.1). We will preserve the indices of $A$ in our examinations of its submatrices; for example, as long $i \in \mathcal{C}_1$ and $j \in \mathcal{C}_2$, $(A_{12})_{ij} = a_{ij}$. The inverse of $I - A_{22}$ (when it exists) is nonnegative, via Lemma 2.8. The matrix

$$S(\mathcal{C}_1) = A_{11} + A_{12} \left(I - A_{22}\right)^{-1} A_{21}$$

is entrywise nonnegative. Every row sum of $A$ is 1 and so we have

$$A_{11}\mathbb{1} + A_{12}\mathbb{1} = \mathbb{1} \quad \text{and} \quad A_{21}\mathbb{1} + A_{22}\mathbb{1} = \mathbb{1};$$

these equalities in turn imply that

$$A_{11}\mathbb{1} = \mathbb{1} - A_{12}\mathbb{1} \quad \text{and} \quad A_{21}\mathbb{1} = \left(I - A_{22}\right)\mathbb{1}.$$

Thus,

$$
\begin{aligned}
S(\mathcal{C}_1)\mathbb{1} &= A_{11}\mathbb{1} + A_{12}\left(I - A_{22}\right)^{-1} A_{21}\mathbb{1} \\
&= \left(\mathbb{1} - A_{12}\mathbb{1}\right) + A_{12}\left(I - A_{22}\right)^{-1}\left(I - A_{22}\right)\mathbb{1} \\
&= \mathbb{1} - A_{12}\mathbb{1} + A_{12}\mathbb{1} \\
&= \mathbb{1}.
\end{aligned}
$$

Now, for $i, j \in \mathcal{C}_1$, let $p_{ij}^{(t)}$ be the sum of the weights of the directed walks $i \rightsquigarrow_{\mathcal{C}_2} j$ with length equal to $t$. If $t = 1$, $p_{ij}^{(1)} = a_{ij}$; if $t \geq 2$,

$$
p_{ij}^{(t)} = \left(P_{12}P_{22}^{t-2}P_{21}\right)_{ij}.
$$

Let $p_{ij}$ be the sum of the weights of the directed walks $i \rightsquigarrow_{\mathcal{C}_2} j$ (of any length). Then, for $i, j \in \mathcal{C}_1$,

$$
\begin{aligned}
p_{ij} &= p_{ij}^{(1)} + p_{ij}^{(2)} + p_{ij}^{(3)} + p_{ij}^{(4)} + \dots \\
&= a_{ij} + (A_{12}A_{21})_{ij} + (A_{12}A_{22}A_{21})_{ij} + (A_{12}A_{22}^2A_{21})_{ij} + \dots \\
&= \left(A_{11} + A_{12}\left(I + A_{22} + A_{22}^2 + \dots\right)A_{21}\right)_{ij} \\
&= \left(A_{11} + A_{12}\left(I - A_{22}\right)^{-1}A_{21}\right)_{ij} \\
&= s_{ij}.
\end{aligned}
$$

∎

**Remark.** The Markov chain described above has a straightforward interpretation (found in [19]). We observe the chains (in the original process) that have have $x_0 \in \mathcal{C}_1$; every time the process leaves $\mathcal{C}_1$ we imagine "fast-forwarding" until we return to $\mathcal{C}_1$, ignoring any time spent in $\mathcal{C}_2$. That is, if a realization of the original Markov chain is given by

$$x_0, x_1, x_2, \dots$$

where $x_0 \in \mathcal{C}_1$, the corresponding realization of the stochastic complement $S(\mathcal{C}_1)$ is obtained by deleting the elements of the sequence contained in $\mathcal{C}_2$; *i.e.* it is

$$x_0, x_{t_1}, x_{t_2}, \ldots$$

where

$$t_1 = \inf\{t \geq 1 : x_t \in \mathcal{C}_1\},\ t_2 = \inf\{t \geq t_1 + 1 : x_t \in \mathcal{C}_1\},\ t_3 = \inf\{t \geq t_2 + 1 : x_t \in \mathcal{C}_1\},$$

and so forth.

### 4.1.1 Stochastic complements of substochastic matrices

Let $B$ be a substochastic matrix with state space $\mathcal{C}$. Recall that an essential class of states with respect to $B$ is a subset $\mathcal{E} \subseteq \mathcal{C}$ such that $B(\mathcal{E})$ is irreducible and stochastic. Such a collection exists if and only if $\rho(B) = 1$, in which case $B$ is not properly substochastic.

We define a stochastic complement of a substochastic matrix in exactly the same manner as for a stochastic matrix. That is, let $B$ be a substochastic matrix with state space $\mathcal{C}$ and let $\{\mathcal{C}_1, \mathcal{C}_2\}$ be a partition of $\mathcal{C}$ such that $\mathcal{C}_2$ does not contain an essential class of states. Express

$$B \cong \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where $B_{ij} = B(\mathcal{C}_i, \mathcal{C}_j)$. Then, we define the stochastic complement of $\mathcal{C}_1$ to be the matrix

$$S(\mathcal{C}_1) = B_{11} + B_{12}\left(I - B_{22}\right)^{-1} B_{21}.$$

The stochastic matrices of order $n$ are a subfamily of the substochastic matrices of order $n$. In the following section, we will prove a number of results concerning stochastic complements of substochastic matrices, with the understanding that these apply to stochastic complements of stochastic matrices.

## 4.2 Properties

Let $A$ be a substochastic matrix and let $\mathcal{S}$ be the associated state space. Let $\{\mathcal{C}_1, \mathcal{C}_2\}$ be a partition of $\mathcal{S}$ such that the complement $S(\mathcal{C}_1)$ exists. We will use the notation

$$A \setminus \mathcal{C}_2 = S(\mathcal{C}_1).$$

*i.e.* The matrix $A \setminus \mathcal{C}$ is the stochastic complement corresponding to the partition of $\mathcal{S}$ into $\mathcal{C}_1 = \mathcal{S} \setminus \mathcal{C}$ and $\mathcal{C}_2 = \mathcal{C}$. If $\mathcal{C} = \{i\}$ (that is, if we are removing a single state) we use the notation $A \setminus i$ to represent the stochastic complement of $\mathcal{S} \setminus i$. We will further define the trivial complement $A \setminus \emptyset = A$.

We will preserve indices between a matrix and its various complements. For example, if $A$ is a $5 \times 5$ stochastic matrix, $A \setminus \{1, 4\}$ is a $3 \times 3$ stochastic matrix whose rows and columns are indexed by the numbers $\{2, 3, 5\}$.

**Proposition 4.4.** *Let $A$ be a substochastic matrix and let $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$ be a partition of the state space into nonempty sets such that the stochastic complement $S(\mathcal{C}_1)$ exists. Then, $S(\mathcal{C}_1)$ can be obtained via two stochastic complements by removing first $\mathcal{C}_2$ and then $\mathcal{C}_3$. That is,*

$$A \setminus (\mathcal{C}_2 \cup \mathcal{C}_3) = (A \setminus \mathcal{C}_2) \setminus \mathcal{C}_3.$$

**Proof** Express $A$ as

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

where $A_{ij} = A(\mathcal{C}_i, \mathcal{C}_j)$. We perform the inverse calculation

$$(I - A(\mathcal{C}_2 \cup \mathcal{C}_3))^{-1} = \begin{bmatrix} I - A_{22} & -A_{23} \\ -A_{32} & I - A_{33} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} X^{-1} + X^{-1}A_{23}Y^{-1}A_{32}X^{-1} & X^{-1}A_{23}Y^{-1} \\ Y^{-1}A_{32}X^{-1} & Y^{-1} \end{bmatrix},$$

where

$$X = I - A_{22} \quad \text{and} \quad Y = I - A_{33} - A_{32}X^{-1}A_{23}.$$

So, the stochastic complement is $S(\mathcal{C}_1) =$

$$A_{11} + \begin{bmatrix} A_{12} & A_{13} \end{bmatrix} \begin{bmatrix} X^{-1} + X^{-1}A_{23}Y^{-1}A_{32}X^{-1} & X^{-1}A_{23}Y^{-1} \\ Y^{-1}A_{32}X^{-1} & Y^{-1} \end{bmatrix} \begin{bmatrix} A_{21} \\ A_{31} \end{bmatrix}$$

$$\begin{aligned} = \; A_{11} \; &+ \; A_{12}X^{-1}A_{21} \\ &+ \; A_{12}X^{-1}A_{23}Y^{-1}A_{32}X^{-1}A_{21} \\ &+ \; A_{12}X^{-1}A_{23}Y^{-1}A_{31} \\ &+ \; A_{13}Y^{-1}A_{32}X^{-1}A_{21} \\ &+ \; A_{13}Y^{-1}A_{31}. \end{aligned}$$

If we first remove $\mathcal{C}_2$ via a stochastic complement we obtain

$$A \setminus \mathcal{C}_2 = \begin{bmatrix} A_{11} & A_{13} \\ A_{31} & A_{33} \end{bmatrix} + \begin{bmatrix} A_{12} \\ A_{32} \end{bmatrix} (I - A_{22})^{-1} \begin{bmatrix} A_{21} & A_{23} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} + A_{12}X^{-1}A_{21} & A_{13} + A_{12}X^{-1}A_{23} \\ A_{31} + A_{32}X^{-1}A_{21} & A_{33} + A_{32}X^{-1}A_{23} \end{bmatrix}.$$

Note that the lower-right diagonal block is $I - Y$. Then, removing $\mathcal{C}_3$ (which corresponds to the lower-right block) obtains $(A \setminus \mathcal{C}_2) \setminus \mathcal{C}_3 =$

$$A_{11} + A_{12}X^{-1}A_{21} + \left(A_{13} + A_{12}X^{-1}A_{23}\right) Y^{-1} \left(A_{31} + A_{32}X^{-1}A_{21}\right);$$

expansion of this expression shows that it is equal to $S(\mathcal{C}_1) = A \setminus (\mathcal{C}_2 \cup \mathcal{C}_3)$. ∎

**Corollary 4.5.** *Let $A$ be a substochastic matrix with state space $\mathcal{S}$ and let $\mathcal{C} \subseteq \mathcal{S}$ be a collection of states that does not contain an entire essential class. Then, the stochastic complement $A \setminus \mathcal{C}$ can be formed by removing the states $i \in \mathcal{C}$, one at a time, via stochastic complements. That is, let $\mathcal{C} = \{i_1, \ldots, i_k\}$, let $A^{(0)} = A$ and for $s = 1, \ldots, k$ let $A^{(s)} = A^{(s-1)} \setminus i_s$. Then, $A \setminus \mathcal{C} = A^{(k)}$.*

Removing a single state via a complement is a simple procedure, computationally. Let $A$ be a stochastic matrix with associated state space $\mathcal{S}$, let $i \in \mathcal{S}$ and let $\mathcal{C} = \mathcal{S} \setminus i$. If we permute $A$ so that

$$A \cong \begin{bmatrix} B & w \\ v^T & a_{ii} \end{bmatrix},$$

where $B = A(\mathcal{C})$, we have

$$S(\mathcal{C}) = A \setminus i = B + \frac{1}{1 - a_{ii}} w v^T,$$

where $v$ and $w$ are column vectors. Note that this complement exists if and only if $a_{ii} \neq 1$.

This corollary is of great utility in applications; determining whether or not $I - A(\mathcal{C})$ is nonsingular and then calculating its inverse can be a somewhat complex computational task. However, it is unnecessary.

Suppose that we want to calculate the stochastic complement $A\backslash\mathcal{C}$, if possible. We simply begin removing the states in $\mathcal{C}$ via complements one at a time. If we discover at some point that an intermediate complement has $s_{ii} = 1$ (where $i$ is the state we intend to remove) we have determined that there is an essential class contained in $\mathcal{C}$. Moreover, we have identified one of its members – the final state we attempted to remove.

Further, removing the states one at a time is essentially no more costly in a computational sense. Suppose that there are $n$ states and we are attempting to calculate the complement which removes a subcollection of $m$ states. Removing all $m$ states at once requires the calculation of an inverse of order $m$, then calculating $(n-m)^2 m$ vector products (the three-fold matrix product in the formula) and then performing a matrix addition. Removing the states one at a time only requires $m$ scalar multiplications, vector products and matrix additions. Analysis shows that the complexities of the two tasks are of the same order.

A stronger version of Proposition 4.6 is presented in [19]. The theorem presented there is used to construct a very interesting algorithm that builds the stationary vector for a stochastic matrix out of the stationary vectors of its stochastic complements.

**Proposition 4.6.** *Let $A$ be a stochastic matrix with state space $\mathcal{S}$ and let $\mathcal{C} \subseteq \mathcal{S}$ be a collection of states such that the stochastic complement $A \setminus \mathcal{C}$ exists. Let $\pi$ be a left eigenvector of $A$ corresponding to the eigenvalue* 1*. Then, $\pi(\mathcal{S} \setminus \mathcal{C})$ is a left eigenvector of $A \setminus \mathcal{C}$ corresponding to the eigenvalue* 1*.*

**Proof** Express $A$ as and $\pi$ as

$$A \cong \begin{bmatrix} A_{11} & A_{12} \\ \\ A_{21} & A_{22} \end{bmatrix} \text{ and } \pi \cong \begin{bmatrix} \pi_1 \\ \\ \pi_2 \end{bmatrix},$$

where $\mathcal{C}_1 = \mathcal{S} \setminus \mathcal{C}$, $\mathcal{C}_2 = \mathcal{C}$, $A_{ij} = A(\mathcal{C}_i, \mathcal{C}_j)$ and $\pi_i = \pi(\mathcal{C}_i)$. We have

$$\pi_1^T A_{11} + \pi_2^T A_{21} = \pi_1^T \text{ and } \pi_1^T A_{12} + \pi_2^T A_{22} = \pi_2^T.$$

Thus,

$$\pi_1^T A_{11} = \pi_1^T - \pi_2^T A_{21} \text{ and } \pi_1^T A_{12} = \pi_2^T \left(I - A_{22}\right).$$

Therefore

$$\begin{aligned} \pi_1^T \left(A \setminus \mathcal{C}\right) &= \pi_1^T A_{11} + \pi_1^T A_{12} \left(I - A_{22}\right)^{-1} A_{21} \\ \\ &= \left(\pi_1^T - \pi_2^T A_{21}\right) + \pi_2^T \left(I - A_{22}\right) \left(I - A_{22}\right)^{-1} A_{21} \\ \\ &= \pi_1^T - \pi_2^T A_{21} + \pi_2^T A_{21} \\ \\ &= \pi_1^T. \end{aligned}$$

Now, we need to show that $\pi_1 \neq 0$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the essential classes of states contained in $\mathcal{S}$. By Corollary 2.4, there is a basis of $m$ eigenvectors,

$$\{\pi_1, \ldots, \pi_m\},$$

for the left eigenspace of $A$ corresponding to the eigenvalue 1 such that for all $i \in \mathcal{S}$, $\pi_k(i) \neq 0$ if and only if $i \in \mathcal{E}_k$. Since the vector $\pi$ is an eigenvector, we have $\pi \neq 0$, and so there is at least one essential class of states $\mathcal{E}_k$ such that $\pi(i) \neq 0$ for all $i \in \mathcal{E}_k$. By Corollary 2.7, $\mathcal{E}_k \nsubseteq \mathcal{C}$, and so we see that $\pi(i) \neq 0$ for at least one state $i$. ∎

**Proposition 4.7.** *Let $A$ be a reversible substochastic matrix and let $A\backslash\mathcal{C}$ be a stochastic complement. Then, $A \setminus \mathcal{C}$ is reversible.*

**Proof** Via Corollary 4.5, it is sufficient to show that if $A$ is a reversible substochastic matrix and $a_{ii} \neq 1$, then $A\backslash i$ is reversible, as well. Let $A$ be a reversible substochastic matrix with associated state space $\mathcal{S}$. Let $i \in \mathcal{S}$ and suppose that $a_{ii} < 1$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Express

$$A \cong \begin{bmatrix} B & w \\ v^T & a_{ii} \end{bmatrix} \quad \text{and} \quad \Pi \cong \begin{bmatrix} \Pi_1 & 0 \\ 0 & \pi_i \end{bmatrix},$$

where $B = A(\mathcal{S} \setminus i)$ and $\Pi_1 = \Pi(\mathcal{S} \setminus i)$ are the principal submatrices on $\mathcal{C} = \mathcal{S} \setminus i$. Then, $\Pi_1 B$ is symmetric and $\Pi_1 w = \pi_i v$. We see that $\Pi_1(A \setminus i)$ is symmetric:

$$
\begin{aligned}
(\Pi_1(A \setminus i))^T &= \left(\Pi_1 B + \tfrac{1}{1-a_{ii}}\Pi_1 w v^T\right)^T \\
&= (\Pi_1 B)^T + \tfrac{1}{1-a_{ii}} v \left(w^T \Pi_1\right) \\
&= \Pi_1 B + \tfrac{1}{1-a_{ii}} \left(\tfrac{1}{\pi_i}\Pi_1 w\right)\left(\pi_i v^T\right) \\
&= \Pi_1 B + \tfrac{1}{1-a_{ii}}\Pi_1 w v^T \\
&= \Pi_1(A \setminus i).
\end{aligned}
$$

∎

We cannot construct a converse of the above theorem without imposing further conditions on $A$; it is possible that every proper stochastic complement of $A$ is reversible but $A$ itself is not reversible. For example, let

$$
A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.
$$

A necessary (but not sufficient) condition for $A$ to be reversible is that $a_{ji} \neq 0$ whenever $a_{ij} \neq 0$; so the above $A$ is not reversible. Yet, any proper stochastic complement of $A$ is equal to

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ or } \begin{bmatrix} 1 \end{bmatrix},$$

both of which are reversible stochastic matrices.

**Proposition 4.8.** *Let $A$ be a stochastic matrix with state space $\mathcal{S}$ and suppose that the stochastic complement $S(\mathcal{C})$ exists. Then,*

1. *for any $i, j \in \mathcal{C}$, we have $i \prec j$ with respect to $A$ if and only if $i \prec j$ with respect to $S(\mathcal{C})$, and*

2. *if $\mathcal{E}$ is an essential class with respect to $A$, then $\mathcal{E} \cap \mathcal{C}$ is an essential class with respect to $S(\mathcal{C})$.*

**Proof** Let $\mathcal{C}_1 = \mathcal{C}$ and $\mathcal{C}_2 = \mathcal{S} \setminus \mathcal{C}$. Let $i, j \in \mathcal{C}_1$ and suppose there is a directed walk $i \rightsquigarrow j$ in the digraph associated with $A$. Such a directed walk can then be expressed as

$$i = i_0 \rightsquigarrow_{\mathcal{C}_2} i_1 \rightsquigarrow_{\mathcal{C}_2} \cdots \rightsquigarrow_{\mathcal{C}_2} i_t = j$$

where each $i_s \in \mathcal{C}_1$. (The vertices $i_s$ in the walk are simply those that are contained in $\mathcal{C}_1$; these are then connected by directed walks through $\mathcal{C}_2$.) Via Proposition 4.3, we have $i_s \prec i_{s+1}$ with respect to $S(\mathcal{C}_1)$ and so $i \prec j$ with respect to $S(\mathcal{C}_1)$. As well, it is clear that if $i \prec j$ under $S(\mathcal{C}_1)$ then $i \prec j$ under $A$. The second statement is a direct consequence of the first. ∎

**Proposition 4.9.** *Let $A$ be a stochastic matrix with state space $\mathcal{S}$ and suppose that the stochastic complement $S(\mathcal{C})$ exists. Then, the Markov chains associated with $A$ and $S(\mathcal{C})$ possess the same number of essential classes, and the multiplicities of* 1 *as an eigenvalue of the matrices $A$ and $S(\mathcal{C})$ are equal.*

**Proof** Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the essential classes with respect to $A$ and let $\mathcal{E}_0$ be the collection of transient states. By Proposition 4.8, each $\mathcal{E}_k \cap \mathcal{C}$ is an essential class with respect to $S(\mathcal{C})$. We simply need show that the states contained in $\mathcal{E}_0 \cap \mathcal{C}$ are transient with respect to $S(\mathcal{C})$.

Let $i \in \mathcal{E}_0 \cap \mathcal{C}$. Since $i \in \mathcal{E}_0$, there is a recurrent state $j$ (with respect to $A$) such that

$$i \prec_A j, \text{ but } j \not\prec_A i.$$

Let $\mathcal{E}_k$ be the essential class that contains $j$ and let $j' \in \mathcal{E}_k \cap \mathcal{C}$. Then, $j'$ is recurrent with respect to $S(\mathcal{C})$,

$$i \prec_{S(\mathcal{C})} j', \text{ and } j' \not\prec_{S(\mathcal{C})} i.$$

The state $i$ must be transient with respect to $S(\mathcal{C})$.

The fact that the multiplicities of 1 as an eigenvalue coincide is then a consequence of Theorem 2.3. ∎

Corollary 4.10 is a consequence of Propositions 4.6 and 4.9.

**Corollary 4.10.** *Let $A$ be an irreducible stochastic matrix with state space $\mathcal{S}$ and unique stationary distribution $\pi$. Let $\mathcal{C} \subseteq \mathcal{S}$ be a nonempty collection of states and let $\hat{\pi} = \pi(\mathcal{S} \setminus \mathcal{C})$. Then, $A \setminus \mathcal{C}$ is irreducible and has unique stationary distribution equal to*

$$\frac{1}{\hat{\pi}^T \mathbb{1}} \hat{\pi}.$$

**Proposition 4.11.** *Let $A$ be a stochastic matrix with state space $\mathcal{S}$ and let $\mathcal{C}_1 \subseteq \mathcal{S}$. Then, the stochastic complement $S(\mathcal{C}_1)$ is equal to the identity matrix if and only if $\mathcal{C}_1$ consists of exactly one member from each essential class. Thus, there exists a collection $\mathcal{C}_2 \subset \mathcal{S}$ such that $A \setminus \mathcal{C}_2 = I_m$ if and only if $\mathcal{S}$ contains exactly $m$ distinct essential classes of states.*

**Proof** Let $\mathcal{C}_1$ consist of one member from each essential class with respect to $A$; let $m$ be the number of essential classes. Then, $S(\mathcal{C}_1)$ has a state space of order $m$ and $m$ distinct essential classes. Clearly, we must have $S(\mathcal{C}_1) = I_m$.

Conversely, if $S(\mathcal{C}_1) = I_m$ then $\mathcal{S}$ contains $m$ essential classes, with respect to $A$ and $\mathcal{C}_1$ contains at least one member from each. However, we have $|\mathcal{C}_1| = m$ and so $\mathcal{C}_1$ contains exactly one member from each essential class. ∎

In Appendix G we summarise many of the important properties of the stochastic complement. We have included this summary for quick reference, as these properties will appear repeatedly in later sections (especially Appendices A and B).

### 4.2.1 Schur complements

Let $M$ be a square complex matrix with index set $\mathcal{S}$ and let $(\mathcal{C}_1, \mathcal{C}_2)$ be a partition of $\mathcal{S}$. Express

$$M \cong \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

where $M_{ij} = M(\mathcal{C}_i, \mathcal{C}_j)$. If the matrix $M_{22}$ is nonsingular, the *Schur complement of* $M_{22}$ *in* $M$ is defined to be the matrix

$$M/M_{22} = M_{11} - M_{12}M_{22}^{-1}M_{21}.$$

An extensive survey of the Schur complement appears in [25].

Suppose that the matrix $M$, as above, is nonsingular and express $M^{-1} = [\tilde{M}_{ij}]$ (that is, $\tilde{M}_{ij} = M^{-1}(\mathcal{C}_i, \mathcal{C}_j)$). If $M_{22}$ is nonsingular, then $\tilde{M}_{11}$ is nonsingular, as well, and

$$M/M_{22} = \tilde{M}_{11}^{-1}.$$

It is important to note that if $M$ is singular, the Schur complement of $M_{22}$ in $M$ may still exist – the above formula is simply a property that holds in the invertible case.

The following well-known formula relates the Schur complement to the determinant of a matrix: let $M$ be a square complex matrix and let $N$ be a nonsingular principal submatrix of $M$; then,

$$\det(M) = \det(N)\det(M/N)$$

([14, Section 0.8.5]).

In [19], the following relationship between the Schur complement and the stochastic complement is noted. Let $A$ be a substochastic matrix on the state space $\mathcal{S}$ and let $\mathcal{C} \subseteq \mathcal{S}$ be such that the stochastic complement $S(\mathcal{C})$ exists. Let $\mathcal{C}_1 = \mathcal{C}$, $\mathcal{C}_2 = \mathcal{S} \setminus \mathcal{C}$ and, for $i, j \in \{1, 2\}$, let $A_{ij} = A(\mathcal{C}_i, \mathcal{C}_j)$. Then,

$$\begin{aligned} S(\mathcal{C}) &= A \setminus \mathcal{C}_2 \\ &= A_{11} + A_{12}(I - A_{22})^{-1}A_{21} \\ &= I - ((I - A_{11}) - A_{12}(I - A_{22})^{-1}A_{21}) \\ &= I - (I - A)/(I - A_{22}). \end{aligned}$$

This is the inspiration for our own notation for the stochastic complement; when $A$ is substochastic and $\mathcal{C}$ does not contain an entire essential class,

$$A \setminus \mathcal{C} = I - (I - A)/(I - A(\mathcal{C})).$$

We have expressed the stochastic complement with set difference notation to emphasise our use of the concept. In this work, the stochastic complement is a method of deleting states from a Markov chain – a stochastic complement $A \setminus \mathcal{C}$ corresponds to a Markov chain obtained removing or "ignoring" the collection of states $\mathcal{C}$.

Many of the properties of the stochastic complement can be derived from known properties of the Schur complement. For example, Proposition 4.4 can be obtained

from the well-known *quotient property* of the Schur complement (discussed in [6]). We have, however, included proofs of every statement in order to achieve a better understanding of these properties.

## 4.3 Convergence of nearly uncoupled Markov chains

We discuss the effects that the property of being nearly uncoupled can have on the convergence of a nearly uncoupled Markov chain to its stationary distribution.

Let $X$ be an irreducible Markov chain on the state space $\mathcal{S}$ with transition matrix $A$. Via the Perron-Frobenius Theorem and Theorem 2.3, 1 is a simple eigenvalue of $A$.

We say that the Markov chain is *periodic* if $\mathcal{S}$ can be partitioned into collections $\mathcal{C}_1, \ldots, \mathcal{C}_p$ such that if $x_t \in \mathcal{C}_k$ then, necessarily, $x_{t+1} \in \mathcal{C}_{k+1}$ (with the convention that $\mathcal{C}_{p+1} = \mathcal{C}_1$). This occurs if and only if for all $k$ and $l$ such that $l \not\equiv k+1$ (modulo $p$), we have $A(\mathcal{C}_k, \mathcal{C}_l) = 0$; that is, if

$$A \cong \begin{bmatrix} 0 & A_{11} & & \\ & \ddots & \ddots & \\ & & \ddots & A_{p-1,p} \\ A_{p1} & & & 0 \end{bmatrix}$$

where the diagonal 0-blocks are square and the unspecified blocks are 0, then the Markov chain associated with $X$ is periodic.

It is known that the following are equivalent, given that $X$ and its transition matrix $A$ are irreducible:

1. The Markov chain $X$ is periodic.

2. The matrix $A$ has an eigenvalue $\lambda \neq 1$ such that $|\lambda| = 1$.

3. There is $k \geq 2$ such that $A^k$ is reducible.

4. There is a nonnegative vector $v$ such that the sequence

$$v^T \ , \ v^T A \ , \ v^T A^2 \ , \ \ldots$$

   fails to converge.

The above set of statements is derived from theorems found in [12, Chapter 8], [14, Chapter 8] and [23, Chapter 4]. We refer to $X$ as *aperiodic* if it is not periodic.

Let $X$ be an aperiodic irreducible Markov chain on the state space $\mathcal{S}$, let $A$ be the associated stochastic matrix and let $v$ be the initial distribution:

$$v_i = \mathbb{P}\left[x_0 = i\right].$$

In contrast to item number 4, above, the fact that $X$ is aperiodic implies that

$$\lim_{t \to \infty} v^T A^t = \pi^T.$$

That is, for sufficiently large $t$,

$$\mathbb{P}_v\left[x_t = i\right] \approx \pi_i = \mathbb{P}_\pi\left[x_t = i\right].$$

This may be used as a method to approximate the stationary distribution of a given stochastic matrix. One chooses an arbitrary nonnegative vector $v = v^{(0)}$ and then iterates $(v^{(t+1)})^T = (v^{(t)})^T A$. When the value $\|v^{(t)} - v^{(t-1)}\|$ becomes sufficiently small, $v^{(t)}$ is an approximation of the stationary distribution (usually utilising the euclidean 2-norm or the $\infty$-norm). However, in general, other methods are typically employed – for example, Gaussian elimination or various factorisations of the matrix $A$.

This method can be used even if the Markov chain is periodic. An irreducible stochastic matrix $A$ with a nonzero diagonal entry must be associated with an aperiodic Markov chain (above we noted that a periodic stochastic matrix is permutation-similar to one where the diagonal blocks are 0). So, if $X$ is periodic and irreducible and has transition matrix $A$, the Markov chain $\tilde{X}$ associated with $\tilde{A} = (1-a)I + aA$, where $0 < a < 1$, is aperiodic and irreducible. Moreover, $\tilde{A}$ has the same stationary distribution as $A$.

If the Markov chain is nearly uncoupled with respect to $\epsilon$, this convergence to the stationary distribution can be very irregular. Suppose that $A$ is an irreducible reversible stochastic matrix and that $A$ is nearly uncoupled with respect to $\epsilon$. Let

$$\lambda_{\text{sup}} = \max_{\lambda \in \sigma(A) \setminus 1} \{|\lambda|\}$$

(where $\sigma(A)$ is the set of eigenvalues of $A$). As in the discussion concerning the Perron cluster approach, we note that $A$ has an eigenvalue $\lambda$ with $1-2\epsilon \leq \lambda < 1$; so, $\lambda_{\text{sup}} \geq 1-2\epsilon$. Let $\pi$ be the stationary distribution of $A$ and let

$$\alpha = \frac{\max\{\sqrt{\pi_i}\}}{\min\{\sqrt{\pi_i}\}}.$$

We note that $\alpha \geq 1$ and, without any further assumptions concerning $A$, there is no upper bound on the value of $\alpha$.

Let $v$ be a nonnegative vector with $v^T\mathbb{1} = 1$ and, for each $t \geq 1$, let $v^{(t)} = (v^T A^t)^T$. In [4, Section 1.5], it is shown that

$$\left\|v^{(t)} - \pi\right\|_2 \leq \lambda_s^t \alpha.$$

This bound on the convergence can be insufficient for practical purposes. We have

$$\lambda_s^t \alpha \geq (1-2\epsilon)^t \alpha;$$

for small values of $\epsilon$, the convergence $(1-2\epsilon)^t \to 0$ is slow (and it is entirely possible that the value $\alpha$ is very large).

In [23, Chapter 4], it is shown that if $A$ is not reversible, a similar bound can be produced. We will focus on the reversible case here because the estimates for $\alpha$ and $\lambda_{\text{sup}}$ in the nonreversible case are more involved and less precise.

When the value $\lambda_{\text{sup}}$, described above, is very small, the Markov chain is referred to as *fast-mixing*. In a fast-mixing Markov chain, the convergence of the sequence

$\{v^T A^t\}$ to the stationary distribution is very rapid. A fast-mixing Markov chain is necessarily aperiodic and cannot be nearly uncoupled, as every eigenvalue $\lambda \neq 1$ of the transition matrix has $|\lambda|$ small.

In the research concerning the Perron cluster approach, the authors assume that even though the Markov chain itself is not fast-mixing, it is fast-mixing within each almost invariant aggregate. (As we noted above, in [7, 8], it is assumed that the number of eigenvalues near to 1 is exactly equal to the number of almost invariant aggregates and that the remaining eigenvalues are bounded in absolute value.)

A behaviour of nearly uncoupled Markov chains that can further complicate the convergence to the stationary distribution is examined in [19]. We summarise only the primary result, referred to as *short-run stabilisation*; see [19] for a full exposition.

Let

$$A = \begin{bmatrix} A_{11} & & A_{ij} \\ & \ddots & \\ & & A_{mm} \end{bmatrix}$$

be an irreducible, reversible and aperiodic stochastic matrix where $\gamma_{A_{kk}} \leq \epsilon \mathbb{1}$ for all $k$; let $\mathcal{S}$ be the associated state space and, for each $k$, let $\mathcal{E}_k$ be the states corresponding to $A_{kk}$. For $k = 1, \ldots, m$, let

$$S_k = S(\mathcal{E}_k) = A \setminus (\mathcal{S} \setminus \mathcal{E}_k)$$

be the stochastic complement of $\mathcal{E}_k$. (That is, $S_k$ is obtained by removing each $i \notin \mathcal{E}_k$

via a stochastic complement.) Let

$$\mu_{\sup} = \max_{1 \leq k \leq m} \max_{\lambda \in \sigma(S_k) \setminus 1} \{|\lambda|\}$$

be the maximum absolute value among the eigenvalues $\lambda \neq 1$ of of the stochastic complements $S_k$. Let

$$\pi^T = \begin{bmatrix} a_1 \hat{\pi}_1^T & \cdots & a_m \hat{\pi}_m^T \end{bmatrix}$$

be the stationary distribution of $A$, expressed so that $a_k \hat{\pi}_k = \pi(\mathcal{E}_k)$ and $a_k = \pi(\mathcal{E}_k)^T \mathbb{1}$. Thus, for all $k$, $\hat{\pi}_k^T \mathbb{1} = 1$. Let $v$ be a nonnegative vector such that $v^T \mathbb{1} = 1$. For $t \geq 1$, let $v^{(t)} = (v^T A^t)^T$; we express

$$v^T = \begin{bmatrix} b_1 v_1^T & \cdots & b_m v_m^T \end{bmatrix} \text{ and } (v^{(t)})^T = \begin{bmatrix} b_1^{(t)} (v_1^{(t)})^T & \cdots & b_m^{(t)} (v_m^{(t)})^T \end{bmatrix},$$

again, so that $v_k^T \mathbb{1} = 1$ and $(v_k^{(t)})^T \mathbb{1} = 1$ for all $k$ and $t$. Finally, let

$$\tilde{\pi}^T = \begin{bmatrix} b_1 \hat{\pi}_1^T & \cdots & b_m \hat{\pi}_m^T \end{bmatrix}$$

We emphasise that $\tilde{\pi} \neq \pi$ (in general).

Now, via our assumptions and previous discussion, we have $v^{(t)} \to \pi$ as $t \to \infty$. Thus, $b_k^{(t)} \to a_k$ and $v_k^{(t)} \to \hat{\pi}_k$ as $t \to \infty$, for all $k$.

The problem that occurs in short-run stabilisation is that the convergence of $v_k^{(t)} \to \hat{\pi}_k$ can be much faster than the convergence of $b_k^{(t)} \to a_k$. This can cause the vector $\tilde{\pi}$ to act as a sort of pseudo-limit for the sequence $\{v^{(t)}\}$, for the initial part of the sequence. Specifically, in [19], it is shown that

$$\left\| v^{(t)} - \tilde{\pi} \right\|_\infty \leq \epsilon t + \alpha \mu_{\text{sup}}^t,$$

and that this bound can be very close to the actual value of $\|v^{(t)} - \tilde{\pi}\|_\infty$. The assumption that $A$ is reversible allows us to use the value

$$\alpha = \frac{\max\{\sqrt{\pi_i}\}}{\min\{\sqrt{\pi_i}\}}.$$

If $A$ is not reversible, a slightly more complicated formula for $\alpha$, which still satisfies $\alpha \geq 1$, is used.

The function

$$f(t) = \epsilon t + \alpha \mu_{\text{sup}}^t$$

is unbounded as $t \to \infty$. However, if $\mu_{\text{sup}}$ is sufficiently small, $f(t)$ is, correspondingly small for the initial tail of the sequence (small values of $t$). Thus, in calculating an estimate of the stationary distribution via the power method above, one may see the values $\|v^{(t)} - v^{(t-1)}\|_\infty$ become very small well before the values $\|v^{(t)} - \pi\|_\infty$ do so.

# Chapter 5

# An algorithm for constructing almost invariant aggregates of a reversible Markov chain

We present the first of a collection of algorithms that attempt to construct almost invariant aggregates in the state space of a reversible Markov chain.

## 5.1 The maximum entry algorithm

Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\mathcal{C} \subseteq \mathcal{S}$ be such that the stochastic complement $\hat{A} = A \setminus \mathcal{C}$ exists. Recall that the state space of $\hat{A}$ is defined to be $\mathcal{S} \setminus \mathcal{C}$. For example, let $\mathcal{S} = \{1, 2, 3, 4, 5\}$ and $\mathcal{C}_1 = \{2, 3\}$. Then, the matrix $\hat{A} = A \setminus \mathcal{C}_1$ is expressed as

$$\hat{A} = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{14} & \hat{a}_{15} \\ \hat{a}_{41} & \hat{a}_{44} & \hat{a}_{45} \\ \hat{a}_{51} & \hat{a}_{54} & \hat{a}_{55} \end{bmatrix}.$$

In this manner, the entries $a_{ij}$ and $\hat{a}_{ij}$ represent the probability of transitioning between the same two states in two different Markov chains. As well, it allows for a simpler expression of successive stochastic complements. For example, let $\mathcal{S}$ and $\mathcal{C}_1$ be as above and let $\mathcal{C}_2 = \{1\}$. Then, the stochastic complement

$$\tilde{A} = A \setminus (\mathcal{C}_1 \cup \mathcal{C}_2)$$

is expressed as

$$\tilde{A} = \begin{bmatrix} \tilde{a}_{44} & \tilde{a}_{45} \\ \tilde{a}_{54} & \tilde{a}_{55} \end{bmatrix}$$

and satisfies $\tilde{A} = \hat{A} \setminus \mathcal{C}_2$.

Typically, mathematical software (such as MatLab) does not include such a structure. The indices of a matrix of order $m$ in storage are forced to be the collection $\{1, \ldots, m\}$ (or, occasionally, $\{0, \ldots, m-1\}$). It is somewhat straightforward to implement the index assignment described above. Let $A$ be a stochastic matrix of order $n$ and let

$$z = \begin{bmatrix} 1 & 2 & \cdots & n \end{bmatrix}.$$

Let $1 \leq i \leq n$ and express

$$A = \begin{bmatrix} A_{11} & v_1 & A_{12} \\ w_1^T & a_{ii} & w_2^T \\ A_{21} & v_2 & A_{22} \end{bmatrix} \text{ and } z = \begin{bmatrix} z_1^T & i & z_2^T \end{bmatrix}.$$

Then, we define

$$\hat{A} = A \setminus i = \begin{bmatrix} A_{11} + \frac{1}{1-a_{ii}} v_1 w_1^T & A_{12} + \frac{1}{1-a_{ii}} v_1 w_2^T \\ A_{21} + \frac{1}{1-a_{ii}} v_2 w_1^T & A_{22} + \frac{1}{1-a_{ii}} v_2 w_2^T \end{bmatrix} \text{ and } \hat{z} = z \setminus i = \begin{bmatrix} z_1^T & z_2^T \end{bmatrix}.$$

Then, the $j$th state of $\hat{A}$ corresponds to the $z_j$th state of $A$. Similarly, if $\mathcal{C} \subseteq \{1, \ldots, n\}$ is some collection of states such that the stochastic complement $\hat{A} = A \setminus \mathcal{C}$ exists, we define

$$\hat{z} = z \setminus \mathcal{C} = [i]_{i \notin \mathcal{C}}.$$

Then, the $k$th index of $\hat{A}$ corresponds to the $\hat{z}_k$th index of $A$ (if we do not alter the order of the indices $z \setminus \mathcal{C}$).

The maximum entry algorithm produces as an output a digraph $G$. We do not examine any specific implementation of the data structure of a digraph. We will simply presume that a digraph is stored as a list of ordered pairs,

$$\{(i_1, j_1), \ldots, (i_r, j_r)\},$$

where the directed arc $i \to j$ is present in $G$ if and only if $(i, j) = (i_s, j_s)$ for some $s \in \{1, \ldots, r\}$. We will explore below how this digraph will assist us in constructing almost invariant aggregates of the Markov chain in question.

As is usual in writing pseudocode, we will use the convention that commands of the form $x := y$ are to be interpreted as

1. calculate $y$, then

2. replace $x$, in storage, with $y$.

For example $x := x + 1$, increases the value of the variable $x$ by 1. If the variable $x$ has not yet been initialised, the first implementation of a command of the form $x := y$ initialises $x$ to be equal to $y$. We will use the symbol $=$ as a Boolean functional; *i.e.* the statement $x = y$ returns **true** if $x$ and $y$ are equal and **false** otherwise.

The inputs of the maximum entry algorithm are a stochastic matrix $A$ on a finite state space $\mathcal{S}$ and a nonnegative value $\delta < 1$.

**Algorithm 3** The maximum entry algorithm

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that, initially, contains no arcs.
$\mathcal{C} := \emptyset$
**while** the order of $B$ is 2 or greater **do**
  Let $i, j \in \mathcal{S} \setminus \mathcal{C}$ be such that $i \neq j$ and $b_{ij} = \max_{j' \neq i'} \{b_{i'j'}\}$.
  **if** $b_{ij} \leq \delta$ **then**
    Exit the **while** loop.
  **else**
    Add the directed arc $i \to j$ to $G$.
    $B := B \setminus i$
    $\mathcal{C} := \mathcal{C} \cup \{i\}$
  **end if**
**end while**
**return** $G$

Let $A$ be a stochastic matrix and suppose that we have applied Algorithm 3 to $A$.

We refer to each execution of the three commands following the **else** statement as an iteration of the algorithm. Let $r$ be the number of iterations of the algorithm before terminating. For $0 \leq s \leq r$, we refer to the matrix $B$, the digraph $G$ and collection $\mathcal{C}$ in storage after $s$ iterations as the stored data after the $s$th iteration. For $s = 0$, the stored data after the $s$th iteration is simply $B = A$, $\mathcal{C} = \emptyset$ and $G$ equal to the empty graph on $\mathcal{S}$. Let $n = |\mathcal{S}|$; we note that after the $s$th iteration, $B$ has order $(n - s) \times (n - s)$, $\mathcal{C}$ contains $s$ states and $G$ contains $s$ directed arcs.

When the algorithm selects the states $i \neq j$ such that $b_{ij}$ is maximal, it is entirely possible that this maximal value may be attained by multiple off-diagonal entries of $A$. The maximum entry algorithm, as presented above, is nondeterministic, in the sense that if the digraphs $G_1$ and $G_2$ are produced by Algorithm 3 with inputs $A_1 \cong A_2$, it is not necessarily true that $G_1 \cong G_2$. We will assume that we have some deterministic method of selecting the maximal off-diagonal entry; however, we will not assume that this maximal off-diagonal value is unique or that the entry selected by the algorithm possesses any other special properties.

**Proposition 5.1.** *Let $A$ be a stochastic matrix of order $n$ and let $\delta < 1$ be non-negative. Then, Algorithm 3 (with inputs $A$ and $\delta$) will terminate after at most $n - 1$ iterations. Moreover, after each iteration, the stored values of $B$ and $\mathcal{C}$ satisfy $B = A \setminus \mathcal{C}$.*

**Proof** The algorithm begins with $B = A$ and each iteration of the algorithm reduces

the order of $B$ by 1; the algorithm terminates (without executing another any further iterations) if the order of $B$ is equal to 1. Thus, we need simply show that an iteration of the algorithm produces a stochastic complement of $A$. We proceed by induction on $s$, where $s$ is the number of iterations completed by the algorithm.

For $s = 0$, we have $B = A$ and $\mathcal{C} = \emptyset$, implying that $B \cong A \setminus \mathcal{C}$.

Let $s \geq 0$, let $B$ and $\mathcal{C}$ be the stored data after $s$ iterations; suppose that $B = A \setminus \mathcal{C}$. Suppose further that the algorithm executes at least one more iteration after the $s$th iteration. We will show that the stored data $B'$ and $\mathcal{C}'$ after the $(s+1)$th iteration satisfies $B' = A \setminus \mathcal{C}'$.

Let $i, j \in \mathcal{S} \setminus \mathcal{C}$ be such that the algorithm identifies $b_{ij}$ as maximal at the start of the $(s+1)$th iteration. So, since the algorithm does not terminate at this point, $b_{ij} > \delta \geq 0$. Thus, the fact that $b_{ii} + b_{ij} \leq 1$ implies that

$$b_{ii} \leq 1 - b_{ij} < 1 - \delta < 1.$$

So, $1 - b_{ii} > 0$ and the stochastic complement

$$B \setminus i = (A \setminus \mathcal{C}) \setminus i = A \setminus (\mathcal{C} \cup \{i\})$$

exists. The stored data after the $(s+1)$th iteration is $B' = B \setminus i$ and $\mathcal{C}' = \mathcal{C} \cup \{i\}$ and so the proposition holds. ∎

Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\delta < 1$ be nonnegative; suppose that we have applied the maximum entry algorithm to $A$. Let $r$ be the total

number of iterations completed by the algorithm before terminating, let $s \leq r$ and let $\mathcal{C}$ be the stored collection after the $s$th iteration completes. In light of Proposition 5.1, we refer to the collection $\mathcal{C}$ as the *states removed* during the first $s$ iterations and the collection $\mathcal{S} \setminus \mathcal{C}$ as the states not yet removed after the $s$th iteration. For $s = r$, we refer to $\mathcal{C}$ as the states removed by the algorithm and $\mathcal{S} \setminus \mathcal{C}$ as the states not removed.

## 5.2 The output of the maximum entry algorithm

A weakly connected component in the digraph $G$ is an induced subgraph $G(\mathcal{E})$ such that either $\mathcal{E} = \{i\}$ where $i$ is an isolated vertex, or

1. any directed arc in $G$ that has at least one of its endpoints contained in $\mathcal{E}$ in fact has both endpoints contained in $\mathcal{E}$, and

2. if we partition $\mathcal{E}$ into any two nonempty disjoint sets $\{\mathcal{E}_1, \mathcal{E}_2\}$, there is at least one directed arc present in $G$ that has one endpoint contained in $\mathcal{E}_1$ and the other endpoint contained in $\mathcal{E}_2$.

Equivalently a weakly connected component containing more than one vertex is an induced subgraph $G(\mathcal{E})$ that satisfies condition 1, above, and is maximal among such induced subgraphs (it is not a proper subgraph of another induced subgraph satisfying 1). See [12, Section 2.6] for a discussion concerning weak connectivity.

We propose that the weakly connected components of $G$, constructed by Algorithm 3 are strong candidates for almost invariant aggregates of the matrix $A$.

The out-degree of a vertex $i$ contained in a directed graph is the number of directed arcs for which $i$ is the initial vertex. Recall that we use the notation $i \prec_G j$ to represent the fact that the digraph $G$ contains a directed walk of length greater than or equal to 1 with initial vertex $i$ and terminal vertex $j$. A digraph is acyclic if it contains no closed directed walks (a walk with initial and terminal vertices identical). That is, $G$ is acyclic if $i \not\prec_G i$ for every vertex $i$.

**Lemma 5.2.** *Let $G$ be an acyclic digraph where every vertex has out-degree equal to* 1 *or* 0*. Then, for every vertex $i$ with out-degree equal to* 1 *there is a unique vertex $j$ with out-degree equal to* 0 *such that $i \prec_G j$. Further, there is a one-to-one correspondence between the weakly connected components of $G$ and the vertices with out-degree* 0*; namely, every weakly connected component contains a unique vertex with out-degree* 0*.*

**Proof** Let $i$ be a vertex in $G$ with out-degree equal to 1. So, $i$ is the initial vertex of at least one directed walk. Given our assumption that $G$ is acyclic, there is an upper limit to the length of directed walks in $G$, as no directed walk can contain the same vertex multiple times. Let

$$\omega = i \to i_1 \to \cdots \to i_l$$

be a directed walk in $G$ with maximal length $l$ among the directed walks with $i$ as an initial vertex. The vertex $i_l$ must have out-degree equal to 0; otherwise we could construct a strictly longer walk. Further, every directed walk with initial vertex $i$ must be a subgraph of $\omega$. (Since every out-degree is one or zero, there is only one possible choice for $i_1$, and then, if $l \geq 2$, only one possible choice for $i_2$, and so forth.) Thus, the vertex $i_l$ in $\omega$ is the unique vertex in $G$ with out-degree 0 such that $i \prec_G j$.

Now, let $\mathcal{E}$ be the vertex set of a weakly connected component in $G$. The collection $\mathcal{E}$ contains at least one vertex, $i$. Either $i$ itself has out-degree 0 or there is $j$ with out-degree 0 such that $i \prec_G j$. When $i \prec_G j$ and $i \in \mathcal{E}$ we must have $j \in \mathcal{E}$; so, $\mathcal{E}$ contains at least one vertex with out-degree 0. Let $j_1, \ldots, j_m$ be the vertices in $\mathcal{E}$ with out-degree 0 and suppose that $m \geq 2$. For $k = 1, \ldots, m$, let

$$\mathcal{E}_k = \{i \in \mathcal{E} : i \preceq_G j_k\}.$$

Since each $i$ with out-degree 1 cannot precede multiple vertices with out-degree 0, these collections partition $\mathcal{E}$. Since $\mathcal{E}$ is weakly connected, $G$ must contain an arc $i \to j$ where $i \in \mathcal{E}_k$, $j \in \mathcal{E}_l$ and $k \neq l$. But this implies that $i \prec_G j_k$ and $i \prec_G j_l$, which is a contradiction. Thus, each weakly connected component in $G$ contains a unique vertex with out-degree 0. ∎

**Lemma 5.3.** *Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\delta < 1$ be nonnegative; suppose that we have applied Algorithm 3 with inputs $A$ and $\delta$. Let $r$ be*

*the number of iterations completed before termination, let* $0 \leq s \leq r$ *and let* $B$, $\mathcal{C}$ *and* $G$ *be the stored data after the* $s$*th iteration. Then,* $G$ *is acyclic and every vertex has out-degree equal to* 1 *or* 0. *Every member of* $\mathcal{C}$ *has out-degree* 1 *and every member of* $\mathcal{S} \setminus \mathcal{C}$ *has out-degree* 0. *Thus, the collection* $\mathcal{S} \setminus \mathcal{C}$ *consists of one member of every weakly connected component of* $G$.

**Proof** Let $n$ be the order of $A$; without loss of generality, we assume that $\mathcal{S} = \{1, \ldots, n\}$. Relabel the states so that state $n$ was removed at the first iteration of Algorithm 3, state $n-1$ was removed second, and so forth. We will show, by induction on $s$, that if the stored data after $s$ iterations is $\mathcal{C}$ and $G$, then

1. $\mathcal{C} = \{n, n-1, \ldots, n-s+1\}$ (if $s = 0$, then $\mathcal{C} = \emptyset$),
2. every member of $\mathcal{C}$ has out-degree 1 in $G$,
3. every member of $\mathcal{S} \setminus \mathcal{C}$ has out-degree 0, and
4. every directed arc $i \to j$ present in $G$ has $i > j$.

(The fourth condition guarantees that $G$ is acyclic.) For $s = 0$, the algorithm has completed no iterations; so, the digraph $G$ contains no arcs and $\mathcal{C} = \emptyset$. The four conditions clearly hold.

Let $1 \leq s \leq r$ and suppose that the statements hold for $s' = s - 1$. We will show their truth for the stored data $\mathcal{C}$ and $G$ after $s$ iterations, as well. Let $\mathcal{C}'$ and $G'$ be

the stored data after $s'$ iterations. State $i = n - s + 1 = n - s'$ is the state removed during the $s$th iteration; so, the directed arc $i \to j$ added to $G'$ to form $G$ must have

$$j \in (\mathcal{S} \setminus \mathcal{C}') \setminus \{i\} = \mathcal{S} \setminus \{n, n-1, \ldots, n-s+1\} = \{1, \ldots, n-s\}.$$

Therefore, the directed $i \to j$ added to $G'$ to form $G$ has $i > j$. This, together with the fact that every directed arc $i' \to j'$ present in $G'$ has $i' > j'$, implies that every directed arc $i' \to j'$ in $G$ has $i' > j'$.

The addition of the directed arc $i \to j$ increases the out-degree of $i$ by one and leaves every other out-degree fixed. We have $\mathcal{C} = \mathcal{C}' \cup \{i\}$ and $i \in \mathcal{S} \setminus \mathcal{C}'$.

Let $i' \in \mathcal{C}$. If $i' = i$, then $i' \in \mathcal{S} \setminus \mathcal{C}'$ implies that $i'$ has out-degree 0 in $G'$ and $i' = i$ further implies that $i'$ has out-degree 1 in $G$. If $i' \neq i$, then $i' \in \mathcal{C}'$ and $i'$ has the same out-degree in $G$ as in $G'$ (namely, 1).

Let $i' \in \mathcal{S} \setminus \mathcal{C}$, then, since $\mathcal{S} \setminus \mathcal{C} \subseteq \mathcal{S} \setminus \mathcal{C}'$ and $i' \neq i$, $i'$ has out-degree 0 in $G'$ and equal out-degree in $G$.

The concluding statement, that $\mathcal{S} \setminus \mathcal{C}$ consists of one member of every weakly connected component of $G$, is a direct consequence of Lemma 5.2. ∎

Let $X$ be a Markov chain with state space $\mathcal{S}$ and transition matrix $A$; suppose that $X$ is nearly uncoupled with respect to $\epsilon$. Recall that an $\epsilon$-uncoupling of $X$ (and $A$) is a partition $\Psi = \{\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0\}$ of $\mathcal{S}$ where

1. $m \geq 2$,

2. for $k \neq 0$, $\mathcal{E}_k$ is a minimal almost invariant aggregate with respect to $\epsilon$,

3. the collection $\mathcal{E}_0$ is allowed to be empty, and

4. when it is nonempty, $\mathcal{E}_0$ does not contain any almost invariant aggregates as subsets.

When $\mathcal{E}_0$ is nonempty we refer to its member states as near transient states. Near transient states are states that are rarely visited by the Markov chain.

Let the digraph $G$ be formed by an application of Algorithm 3 to $A$; let the arc $i \to j$ be present in $G$. Then, the maximum entry algorithm, at some iteration, constructed a stochastic complement $A \setminus \mathcal{C}$ such that the $ij$th transition probability was maximal. That is, there is a collection $\mathcal{C}' \subseteq \mathcal{S}$ (namely $\mathcal{C}' = \mathcal{S} \setminus \mathcal{C}$) such that whenever the Markov chain visits state $i$, the member of $\mathcal{C}'$ that it is most likely to visit next is $j$. Suppose that $i \in \mathcal{E}$ where $\mathcal{E}$ is an almost invariant aggregate. It seems reasonable to conclude that either $\mathcal{E} \cap \mathcal{C}' = \{i\}$ or that $j \in \mathcal{E}$ (the state most likely to be visited after visiting $i$ should be a member of the almost invariant aggregate containing $i$).

The above reasoning suggests the following conclusion: if $i \in \mathcal{E}_k$ for some $k \neq 0$, then $j \in \mathcal{E}_k$, as well (simply because transitions that exit an almost invariant aggregate are relatively rare). Thus, we suspect that for every arc $i \to j$, present in $G$, either $i \in \mathcal{E}_0$ or $i, j \in \mathcal{E}_k$ for some $k \neq 0$.

Moreover, let $j_1, \ldots, j_m$ be the states not removed by the algorithm. Then, during its final iteration, the maximum entry algorithm constructs the stochastic complement

$$B = S(\{j_1, \ldots, j_m\})$$

and the off-diagonal entries of this matrix are found to be each less than or equal to $\delta$. Thus, if $\delta$ is well-chosen, we suspect that any two of these states are not contained in the same almost invariant aggregate, as transitions between these states seem to have a small chance of occurring.

**Proposition 5.4.** *Let $X$ be a nearly uncoupled Markov chain on the state space $\mathcal{S}$ and let $\Psi = \{\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0\}$ be an $\epsilon$-uncoupling of $X$. Let $G$ be an acyclic digraph with vertex set $\mathcal{S}$ and suppose that*

1. *every vertex in $G$ has out-degree equal to $1$ or $0$,*
2. *for every directed arc $i \to j$ present in $G$, either $i, j \in \mathcal{E}_k$ where $k \neq 0$, or $i \in \mathcal{E}_0$, and*
3. *if $i$ and $j$ are states with out-degree $0$ in $G$, then there is no $\mathcal{E}_k$ where $k \neq 0$ that contains both $i$ and $j$.*

*Let $\mathcal{E}$ be the vertex set of a weakly connected component of $G$ and let $j \in \mathcal{E}$ be the unique member of $\mathcal{E}$ with out-degree equal to $0$. Then, either $\mathcal{E} \subseteq \mathcal{E}_0$ or*

$$\mathcal{E}_k \subseteq \mathcal{E} \subseteq \mathcal{E}_k \cup \mathcal{E}_0 \text{ and } j \in \mathcal{E}_k$$

*for a unique* $k \neq 0$.

**Proof** Let $\mathcal{E}$ be the vertex set of a weakly connected component of $G$. We will show that if $\mathcal{E} \nsubseteq \mathcal{E}_0$, then there is a unique $k \neq 0$ such that

$$\mathcal{E}_k \subseteq \mathcal{E} \subseteq \mathcal{E}_k \cup \mathcal{E}_0.$$

Suppose that $\mathcal{E} \nsubseteq \mathcal{E}_0$. Since $\Psi$ forms a partition of $\mathcal{S}$, there is $k \neq 0$ such that $\mathcal{E} \cap \mathcal{E}_k$ is nonempty; let $i \in \mathcal{E} \cap \mathcal{E}_k$. As well, let $j$ be the unique member of $\mathcal{E}$ with out-degree equal to 0 (such a $j$ exists via Lemma 5.2).

Whenever the arc $i' \to j'$ is present in $G$, we have either $i', j' \in \mathcal{E}_{k'}$ for some $k' \neq 0$ or $i' \in \mathcal{E}_0$; this implies that whenever $i' \prec_G j'$ then either $i', j' \in \mathcal{E}_{k'}$ for some $k' \neq 0$ or $i' \in \mathcal{E}_0$.

So, we have $i \prec_G j$ (again, via Lemma 5.2) and $i \in \mathcal{E}_k$ where $k \neq 0$; thus, $j \in \mathcal{E}_k$. For every other member $i' \in \mathcal{E}$, we have $i' \prec_G j$ and so every member of $\mathcal{E}$ is contained in either $\mathcal{E}_k$ or $\mathcal{E}_0$. That is,

$$\mathcal{E} \subseteq \mathcal{E}_k \cup \mathcal{E}_0.$$

Now, suppose that there is another weakly connected component $\mathcal{E}'$ of $G$, distinct from $\mathcal{E}$, such that $\mathcal{E}' \cap \mathcal{E}_k$ is nonempty. Then, as above, the unique member $j' \in \mathcal{E}'$ with out-degree equal to 0 is contained in $\mathcal{E}_k$. So, this supposition implies that $\mathcal{E}_k$ contains two states, $j$ and $j'$, with out-degree equal to 0. This contradicts the third assumption in the statement and so the weakly connected component $\mathcal{E}$ is the only

weakly connected component that intersects $\mathcal{E}_k$. The weakly connected components of a digraph form a partition of its vertices; so, since $\mathcal{E}_k$ has a nonempty intersection with only one weakly connected component, $\mathcal{E}$, it must be that $\mathcal{E}_k \subseteq \mathcal{E}$.

So, we have shown that for each weakly connected component $\mathcal{E}$, there is $k \neq 0$ such that

$$\mathcal{E}_k \subseteq \mathcal{E} \subseteq \mathcal{E}_k \cup \mathcal{E}_0.$$

Since the collections $\mathcal{E}_k$ are disjoint, this must be satisfied for exactly one almost invariant aggregate $\mathcal{E}_k$. ∎

We emphasise that we are unable to show, in general, that the output of the Maximum entry algorithm satisfies the assumptions of Proposition 5.4. Below, we show that this holds for $\epsilon = 0$; for positive values of $\epsilon$ it is straightforward to construct examples that "fool" the algorithm (see, for example, Appendix D). However, it seems entirely reasonable to assume that the assumptions hold for most of the arcs and states in such an output. Thus, we suspect that, when the input $\delta$ is well-chosen, the weakly connected components of the output of Algorithm 3 consist largely of states from one almost invariant aggregate together with some collection of near-transient states. Experiments (see Appendix C) seem to reinforce this supposition.

### 5.2.1 Direct calculation of the output aggregates

We have structured our algorithm to output a digraph as this digraph then contains interesting topological information about the related Markov chain. Whenever there is a directed path $i \rightsquigarrow j$ of small length in the output digraph, we suspect that the states $i$ and $j$ are closely linked in the associated Markov chain. In Appendix C we explore an idea we refer to as recursive subaggregating which attempts to take advantage of this information to produce a stronger output. However, it may be that the only information of interest to the user is the vertex sets of the weakly connected components. It is very simple to alter our pseudocode so that this simpler output is produced.

We replace the initialisation of the digraph $G$ with the following command

$$\text{Let } \mathcal{B} = (\{1\}, \{2\}, \ldots, \{n\}) .$$

Each of the commands to add a directed arc $i \to j$ to $G$ is then replaced with the command

$$\mathcal{B}_j := \mathcal{B}_j \cup \mathcal{B}_i.$$

Via induction, we can see that after each iteration of the algorithm,

$$\mathcal{B}_j = \{i : i \preceq_G j\}$$

(if we had been constructing the digraph $G$ as usual).

Upon termination, the algorithm returns the collection $\{B_j\}_{j \notin \mathcal{C}}$, where $\mathcal{C}$ is the collection of states removed, via stochastic complements. By Lemma 5.3, these are the vertex sets of the weakly connected components, had we used the original implementation.

## 5.3 Near transient states

We explore the effect that the presence of near transient states can have on the algorithms which utilise the stochastic complement.

**Proposition 5.5.** *Let $A$ be a nearly uncoupled stochastic matrix with respect to $\epsilon$ and let $\Psi = \{\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0\}$ be an $\epsilon$-uncoupling. Let $k \neq 0$ and let $\mathcal{C}$ be a collection of states disjoint from $\mathcal{E}_k$ such that the stochastic complement $\hat{A} = A \setminus \mathcal{C}$ exists. Then, $\mathcal{E}_k$ is an almost invariant aggregate of the Markov chain associated with $\hat{A}$. Thus, if $\mathcal{C} \subseteq \mathcal{E}_0$, then for all $k \neq 0$, $\mathcal{E}_k$ is an almost invariant aggregate of the Markov chain associated with $\hat{A}$.*

**Proof** Let $k \neq 0$ and express

$$A \cong \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

where the first position corresponds to $\mathcal{E}_k$, the second to $\mathcal{C}$ and the third to the

remainder of the state space. The principal submatrix of $\hat{A} = A \setminus \mathcal{C}$ corresponding to $\mathcal{E}_k$ is

$$\hat{A}(\mathcal{E}_k) = A_{11} + A_{12}(I - A_{22})^{-1}A_{21}.$$

The necessary and sufficient condition for $\mathcal{E}_k$ to be an almost invariant aggregate is that the sum of the entries in each row of the principal submatrix corresponding to $\mathcal{E}_k$ is at least $1 - \epsilon$. Thus, since $\hat{A}(\mathcal{E}_k) \geq A_{11}$, if $A_{11}\mathbb{1} \geq (1 - \epsilon)\mathbb{1}$ then we also have $\hat{A}(\mathcal{E}_k)\mathbb{1} \geq (1 - \epsilon)\mathbb{1}$. ∎

Thus, we consider near transient states to be "safe" to remove, in that their removal, via stochastic complements does not affect the basic uncoupled structure of the stochastic matrix involved.

Furthermore, the addition of near transient states to an almost invariant aggregate does not, in general, alter the fact that transitions into and out of that collection rarely occur.

For example, consider the reversible stochastic matrix

$$A = \begin{bmatrix} 1-\epsilon & 0 & \epsilon \\ 0 & 1-\epsilon^2 & \epsilon^2 \\ 1-\epsilon & \epsilon & 0 \end{bmatrix}.$$

The unique stationary distribution of $A$ is

$$\pi = \frac{1}{2}\begin{bmatrix} 1-\epsilon & 1 & \epsilon \end{bmatrix}^T.$$

The unique $\epsilon$-uncoupling of $A$ is into the almost invariant aggregates $\mathcal{E}_1 = \{1\}$ and $\mathcal{E}_2 = \{2\}$ and near transient collection $\mathcal{E}_0 = \{3\}$. The $\pi$-coupling measures of $\mathcal{E}_1$ and $\mathcal{E}_2$ are $1 - \epsilon$ and $1 - \epsilon^2$, respectively.

We note that $a_{31} > a_{32}$ and $a_{13} > a_{23}$ – if state 3 is to be associated with either of states 1 or 2 it seems that it should be with state 1. However, consider the $\pi$-coupling measures associated with the collections $\mathcal{C}_1 = \{1, 3\}$ and $\mathcal{C}_2 = \{2, 3\}$:

$$w_\pi(\mathcal{C}_1) = \frac{\pi_1(a_{11} + a_{13}) + \pi_3(a_{31} + a_{33})}{\pi_1 + \pi_3} = 1 - \epsilon^2$$

$$\text{and } w_\pi(\mathcal{C}_2) = \frac{\pi_2(a_{22} + a_{23}) + \pi_3(a_{32} + a_{33})}{\pi_2 + \pi_3} = \frac{1 + \epsilon^2}{1 + \epsilon} = 1 - \epsilon + \frac{2\epsilon}{1 + \epsilon}.$$

So, adding the near transient state 3 to the almost invariant aggregate $\mathcal{E}_1$ increases the associated $\pi$-coupling measure by $\epsilon - \epsilon^2$, creating a slightly stronger almost invariant aggregate. Adding state 3 to the almost invariant aggregate $\mathcal{E}_2$ increases the $\pi$-coupling measure by

$$-\epsilon + \epsilon^2 + \frac{2\epsilon}{1 + \epsilon} = \frac{\epsilon + \epsilon^3}{1 + \epsilon},$$

again slightly strengthening the almost invariant property. The difference between these two slight increases is, itself, insignificant:

$$\left| \frac{\epsilon + \epsilon^3}{1 + \epsilon} - (\epsilon - \epsilon^2) \right| = \frac{2\epsilon^3}{1 + \epsilon}.$$

If the near-transient state 3 is added to either of the almost invariant aggregates it does not alter the fact that they are almost invariant aggregates, with respect to the

$\pi$-coupling measure, and it makes very little difference which aggregate it is added to.

So, to a certain extent, we are unconcerned with the assignment of near-transient states to aggregates by the maximum entry algorithm. As long as members of distinct almost invariant aggregates are correctly assigned to different weakly connected components of the digraph, the long-term predictive power of the produced aggregates will still be accurate.

## 5.4 A note concerning Appendices A and B

In Appendix A we prove the following proposition.

**Proposition 5.6.** *Let $B$ be an irreducible reversible substochastic matrix of order $m$ such that $\gamma_B \leq \epsilon\mathbb{1}$. Let $\Pi$ be a positive diagonal matrix such that $\Pi B$ is symmetric and let $i$ be such that*

$$\pi_i = \max\{\pi_j\}.$$

*Then,*

$$[\alpha] = B \setminus \{j : j \neq i\}$$

*satisfies*

$$\alpha \geq \frac{(1-\epsilon)^2}{1+(m-2)\epsilon}.$$

Now, suppose that $A$ is a reversible stochastic matrix on the state space $\mathcal{S}$ and suppose that $\mathcal{E} \subseteq \mathcal{S}$ is an almost invariant aggregate with respect to $\epsilon$. Let $\mathcal{C} \subseteq \mathcal{S}$ be such that $\tilde{A} = A \setminus \mathcal{C}$ is well defined and such that $\mathcal{E} \setminus \mathcal{C}$ contains only one state, say $\mathcal{E} \setminus \mathcal{C} = \{i\}$. Let $B = A(\mathcal{E})$ and consider the following calculation of $\tilde{a}_{ii}$. Express

$$A \cong \begin{bmatrix} a_{ii} & v_1^T & v_2^T & v_3^T \\ w_1 & A_{11} & A_{12} & A_{13} \\ w_2 & A_{21} & A_{22} & A_{23} \\ w_3 & A_{31} & A_{32} & A_{33} \end{bmatrix}$$

where the first position corresponds to $i$, the second to

$$\mathcal{C}_1 = \mathcal{C} \cap \mathcal{E} = \{j \in \mathcal{E} : j \neq i\},$$

the third to $\mathcal{C}_2 = \mathcal{C} \setminus \mathcal{E}$ and the fourth to the remainder of the state space. We note that

$$B = A(\mathcal{E}) \cong \begin{bmatrix} a_{ii} & v_1^T \\ w_1 & A_{11} \end{bmatrix} \text{ and } B \setminus \mathcal{C}_1 = [a_{ii} + v_1^T (I - A_{11})^{-1} w_1].$$

We form the stochastic complement $A \setminus \mathcal{C}$ by first removing $\mathcal{C}_1$ and then removing $\mathcal{C}_2$ (Proposition 4.4). We calculate

$$A \setminus \mathcal{C}_1 \cong \begin{bmatrix} a_{ii} & v_2^T & v_3^T \\ w_2 & A_{22} & A_{23} \\ w_3 & A_{32} & A_{33} \end{bmatrix} + \begin{bmatrix} v_1^T \\ A_{21} \\ A_{31} \end{bmatrix} (I - A_{11})^{-1} \begin{bmatrix} w_1 & A_{12} & A_{13} \end{bmatrix}$$

$$= \begin{bmatrix} a_{ii} + v_1^T (I - A_{11})^{-1} w_1 & \tilde{v}_2^T & \tilde{v}_3^T \\ \tilde{w}_2 & \tilde{A}_{22} & \tilde{A}_{23} \\ \tilde{w}_3 & \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix} = \begin{bmatrix} \alpha & \tilde{v}_2^T & \tilde{v}_3^T \\ \tilde{w}_2 & \tilde{A}_{22} & \tilde{A}_{23} \\ \tilde{w}_3 & \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix},$$

where

$$[\alpha] = B \setminus \mathcal{C}_1 = B \setminus \{j \in \mathcal{E} : j \neq i\}.$$

Thus, when we form the stochastic complement $\hat{A} = A \setminus \mathcal{C} = \tilde{A} \setminus \mathcal{C}_2$, the $ii$th entry of $\hat{A}$ is

$$\hat{a}_{ii} = \alpha + \tilde{v}_2^T (I - \tilde{A}_{22})^{-1} \tilde{w}_2 \geq \alpha.$$

So, our above proposition provides a lower bound for the $ii$th entry of $A \setminus \mathcal{C}$, in the case that $i$ is a member of $\mathcal{E}$ with maximal value (among members of $\mathcal{E}$) in the stationary distribution of $A$:

$$\hat{a}_{ii} \geq \frac{(1-\epsilon)^2}{1 + (m-2)\epsilon},$$

where $m = |\mathcal{E}|$.

We will make extensive use of this lower bound in later sections. In the next section, we use this lower bound to show that our Maximum Entry Algorithm avoids a particular type of error, if the input value $\delta$ is well-chosen. Later in this chapter, we make use of this lower bound to produce the Modified Maximum Entry Algorithm

(Algorithm 4). The Modified Maximum Entry Algorithm attempts to avoid errors without having to fine-tune the input value $\delta$.

In Appendix B we attempt to find a similar lower bound, concerning the nonreversible case. We prove the following proposition.

**Proposition 5.7.** *Let $B$ be an irreducible substochastic matrix on the state space $\mathcal{E}$, of order $m$, such that $\gamma_B \leq \epsilon\mathbb{1}$. For each $i \in \mathcal{E}$, let*

$$B \setminus \{j : j \neq i\} = [\alpha(i)].$$

*Then, there is a positive sequence $\beta(j)$ on $\mathcal{E}$ such that*

$$\sum_{j\in\mathcal{E}} \beta(j) = 1 \text{ and } \sum_{j\in\mathcal{E}} \alpha(j)\beta(j) \geq (1-\epsilon)^m.$$

We are unable to characterise those states (in the above proposition) that have $\alpha(i) \geq (1-\epsilon)^m$; however, there must be at least one and, on average, the states satisfy this inequality.

We make use of this lower bound to construct an algorithm for use on nonreversible Markov chains – the Minimum Column Algorithm (Algorithm 10 in Chapter 6).

We have placed the calculations of these lower bounds in the appendices as the proofs are somewhat involved.

## 5.5 The removal of an almost invariant aggregate

There are a number of errors that the maximum entry algorithm may make in decoupling a particular matrix $A$. For example, the digraph $G$ may contain an arc $i \to j$ where $i$ and $j$ belong to disjoint almost invariant aggregates. We show that if the input value $\delta$ is well-chosen, there is one particular type of error that the algorithm will not make.

**Lemma 5.8.** *Let $0 \leq \epsilon < 1$. The function*

$$f(m) = \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon},$$

*where $m$ is a positive integer, is increasing in $m$. We further note that $0 \leq f(m)$, $f(m) \leq m\epsilon$ and that $f(m) < 1$, for $m \geq 1$.*

**Proof** We simply take the derivative of the function $f(z)$:

$$\frac{df}{dz} = \frac{\epsilon\left(1+(z-2)\epsilon\right) - \epsilon^2(z-\epsilon)}{\left(1+(z-2)\epsilon\right)^2} = \frac{\epsilon(1-\epsilon)}{\left(1+(z-2)\epsilon\right)^2} \geq 0.$$

The assumption that $m \geq 1$ implies that $f(m)$ is increasing in $m$.

If $\epsilon = 0$, then $f(m) = 0$, and the three inequalities hold. So, suppose that $0 < \epsilon < 1$. Then, $z \geq 1$ implies that

$$\frac{df}{dz} > 0$$

(since $\epsilon(1-\epsilon) > 0$). We have

$$f(1) = \frac{\epsilon(1-\epsilon)}{1-\epsilon} = \epsilon > 0$$

and

$$\lim_{z\to\infty} \frac{\epsilon(z-\epsilon)}{1+(z-2)\epsilon} = \lim_{z\to\infty} \frac{\epsilon z - \epsilon^2}{\epsilon z + 1 - 2\epsilon} = 1.$$

Thus, for $0 < \epsilon < 1$ and $m \geq 1$, $f(m)$ is strictly increasing in $m$, strictly bounded below by 0 and bounded above by 1, which in turn imply that $0 < f(m) < 1$.

Now, we show that if $m \geq 1$ and $0 \leq \epsilon < 1$, then

$$m\epsilon \geq \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon}.$$

First, we consider the case $m = 1$. Then,

$$\frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon} = \frac{\epsilon(1-\epsilon)}{1-\epsilon} = \epsilon = m\epsilon.$$

Next, suppose that $m \geq 2$. Then, $0 < m - \epsilon \leq m$ and $1 \leq 1 + (m-2)\epsilon$. So,

$$\frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon} \leq \epsilon(m-\epsilon) \leq \epsilon m.$$

∎

**Lemma 5.9.** *Let $A$ be a reversible stochastic matrix on the state space $\mathcal{S}$ and suppose that the digraph $G$ was formed by an application of Algorithm 3. Let $\pi$ be a stationary distribution of $A$. For any $i, j \in \mathcal{S}$, if $i \prec_G j$ then $\pi_i \leq \pi_j$.*

**Proof** We show that for any directed arc $i \to j$ present in $G$, we have $\pi_i \leq \pi_j$. This will clearly imply that if $i \prec_G j$ then $\pi_i \leq \pi_j$.

Suppose that the directed arc $i \to j$ was added to $G$ during the $s$th iteration of the algorithm. Let $\mathcal{C}$ be the states removed during the first $s-1$ iterations and let $B \cong A \setminus \mathcal{C}$ be the stored stochastic complement after the $(s-1)$th iteration. Since the arc $i \to j$ was added at iteration $s$, $b_{ij}$ is maximal among the off-diagonal entries of $B$ and is not equal to 0. In particular, $b_{ij} \geq b_{ji}$; as well, we note that since $B$ is reversible, we have $b_{ji} > 0$, thus $b_{ij} \geq b_{ji} > 0$. Via Proposition 2.12 and Corollary 4.10,

$$\pi_i b_{ij} = \pi_j b_{ji}.$$

Thus, we either have $\pi_i = \pi_j = 0$ or

$$\pi_i = \pi_j \frac{b_{ji}}{b_{ij}} \leq \pi_j.$$

∎

**Proposition 5.10.** *Let $A$ be a nearly uncoupled reversible stochastic matrix with respect to $\epsilon$ on the state space $\mathcal{S}$. Let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate containing $m$ states. If*

$$\delta \geq \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon},$$

*then the maximum entry algorithm (with input $\delta$) will not remove every member of $\mathcal{E}$.*

**Proof** First, suppose that $m = 1$. Then $\mathcal{E} = \{i'\}$ where $a_{i'i'} \geq 1 - \epsilon$. Let $\hat{A} = A \setminus \mathcal{C}$ be a stochastic complement where $i' \notin \mathcal{C}$; then, $\hat{a}_{i'i'} \geq a_{i'i'} \geq 1 - \epsilon$. As well, we have

$$\delta \geq \frac{\epsilon(m - \epsilon)}{1 + (m - 2)\epsilon} = \frac{\epsilon(1 - \epsilon)}{1 - \epsilon} = \epsilon.$$

Now, suppose that the algorithm has completed $s$ iterations, forming the stochastic complement $\hat{A}$ and suppose further that it has not yet removed $i'$ (we may have $s = 0$ and $\hat{A} = A$). The maximum entry algorithm will remove $i'$ at the $(s + 1)$th iteration only if there is some $j' \neq i'$, not yet removed, such that

$$\hat{a}_{i'j'} = \max_{i \neq j} \{\hat{a}_{ij}\}$$

and $\hat{a}_{i'j'} > \delta$. However, the fact that $\hat{a}_{i'i'} \geq 1 - \epsilon \geq 1 - \delta$ implies that for all $j \neq i'$, we have

$$\hat{a}_{i'j} \leq 1 - \hat{a}_{i'i'} \leq \delta.$$

So, no such $j'$ can exist. Therefore, if $m = 1$, the single member of $\mathcal{E}$ will not be removed during any iteration of the algorithm.

We next assume that $m \geq 2$. Since $A$ is reversible, there is a positive diagonal matrix $\Pi$ such that $\Pi A$ is symmetric. Let

$$p = \max_{i \in \mathcal{E}} \{\pi_i\}$$

and let

$$\mathcal{E}_{\max} = \{i \in \mathcal{E} : \pi_i = p\}.$$

Let the digraph $G$ be formed by an application of Algorithm 3 with input

$$\delta \geq \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon},$$

where $m = |\mathcal{E}|$.

We will show that any directed arc present in $G$ that has its initial vertex contained in $\mathcal{E}_{\max}$ must also have its terminal vertex contained in $\mathcal{E}_{\max}$. Since $G$ is acyclic, this will imply that there is at least one member of $\mathcal{E}_{\max}$ that has out-degree 0 in $G$. By Lemma 5.3, such a state will be a member of $\mathcal{E}$ that was not removed by the algorithm.

Suppose that the directed arc $i' \to j'$ was added to $G$ during the $s$th iteration of the while loop of Algorithm 3 and that $i' \in \mathcal{E}_{\max}$. Via Proposition 2.12, Corollary 4.10 and Lemma 5.9, we have $\pi_{i'} \leq \pi_{j'}$. Thus, if $j' \in \mathcal{E}$, then $j' \in \mathcal{E}_{\max}$; so, we will merely need to show that $j' \in \mathcal{E}$.

Let $\mathcal{C}$ be the collection of states removed during the first $s-1$ iterations of the while loop (if $s = 1$, $\mathcal{C} = \emptyset$). If $\mathcal{E} \cup \mathcal{C} = \mathcal{S}$, that is, if every member of $\mathcal{S}$ not contained in $\mathcal{E}$ was removed during the first $s-1$ iterations, then we must have $j' \in \mathcal{E}$. So, we assume that $\mathcal{S}$ contains one or more states contained in neither $\mathcal{E}$ nor $\mathcal{C}$.

Let $\mathcal{C}_{\mathcal{E}} = \mathcal{C} \cap \mathcal{E}$ and let $\mathcal{C}_{\tilde{\mathcal{E}}} = \mathcal{C} \setminus \mathcal{E}$. Let $\hat{A} = A \setminus \mathcal{C}$ be the stored stochastic complement during the $k$th iteration and let $\tilde{A} = A \setminus \mathcal{C}_{\tilde{\mathcal{E}}}$. We note that $\hat{A} = \tilde{A} \setminus \mathcal{C}_{\mathcal{E}}$.

We emphasise, at this point, that either (or both) of the collections $\mathcal{C}_{\mathcal{E}}$ or $\mathcal{C}_{\tilde{\mathcal{E}}}$ may be empty. Thus, we may have $\tilde{A} = A$ and/or $\hat{A} = \tilde{A}$.

As in the proof of Proposition 5.5, we have $\tilde{A}(\mathcal{E}) \geq A(\mathcal{E})$. Thus, since $\mathcal{E}$ is an almost invariant aggregate with respect to $\epsilon$, we have

$$\tilde{A}(\mathcal{E})\mathbb{1} \geq A(\mathcal{E})\mathbb{1} \geq (1-\epsilon)\mathbb{1}.$$

We now consider two cases.

*Case one.* The collection $\mathcal{C}_{\mathcal{E}}$ is empty.

In this case, we have $\hat{A} = \tilde{A}$. Thus, the state $i'$ is the first member of $\mathcal{E}$ to be removed. As noted above, we have

$$\hat{A}(\mathcal{E})\mathbb{1} = \tilde{A}(\mathcal{E})\mathbb{1} \geq (1-\epsilon)\mathbb{1}.$$

Therefore, for all $j \notin \mathcal{E}$ not yet removed by the algorithm, $\hat{a}_{i'j} \leq \epsilon$. Now,

$$\delta \geq \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon} \geq \frac{\epsilon(1-\epsilon)}{1-\epsilon} = \epsilon$$

(via Lemma 5.8 and the fact that $m > 1$). The fact that the directed arc $i' \to j'$ was added at the $s$th iteration implies that $\hat{a}_{i'j'} > \delta$; so, we must have $j' \in \mathcal{E}$.

*Case two.* The collection $\mathcal{C}_{\mathcal{E}}$ is nonempty.

We will make use of Proposition A.14. This proposition is part of a set of proofs found in Appendix A.

Let $m' = |\mathcal{C}_{\mathcal{E}}|$ and note that $m' \leq m-1$. Now, the principal submatrix of $\hat{A}$ corresponding to $\mathcal{E} \setminus \mathcal{C}$ is

$$\hat{A}(\mathcal{E} \setminus \mathcal{C}_A) = \tilde{A}(\mathcal{E}) \setminus \mathcal{C}_{\mathcal{E}}.$$

Since $i' \in \mathcal{E}_{\max}$, every state $j \in \mathcal{C}_{\mathcal{E}}$ has $\pi_j \leq \pi_{i'}$. By Proposition 2.12 and Corollary 4.10, left-multiplying $\tilde{A}(\mathcal{E})$ by the positive diagonal matrix $\Pi(\mathcal{E})$ produces a symmetric matrix. These facts, together with Proposition A.14, imply that the sum of the entries in the row of

$$\tilde{A}(\mathcal{E}) \setminus \mathcal{C}_{\mathcal{E}}$$

corresponding to $i'$ satisfies

$$\sum_{j \in \mathcal{E} \setminus \mathcal{C}_{\mathcal{E}}} \hat{a}_{i'j} \geq \frac{(1-\epsilon)^2}{1+(m'-1)\epsilon}.$$

Since $m' \leq m-1$,

$$\frac{(1-\epsilon)^2}{1+(m'-1)\epsilon} \geq \frac{(1-\epsilon)^2}{1+(m-2)\epsilon}.$$

If we suppose that $j' \notin \mathcal{E}$, then we must have

$$\begin{aligned} \hat{a}_{i'j'} &\leq 1 - \sum_{j \in \mathcal{E} \setminus \mathcal{C}} \hat{a}_{i'j} \\ &\leq 1 - \tfrac{(1-\epsilon)^2}{1+(m-2)\epsilon} \\ &= \tfrac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon} \\ &\leq \delta. \end{aligned}$$

As in case one, the assumption that the directed arc $i' \to j'$ was added during the $s$th iteration implies that $\hat{a}_{i'j'} > \delta$. This contradicts the above conclusion – so, we must have $j' \in \mathcal{E}$. ∎

**Corollary 5.11.** *Let $A$ be a reversible stochastic matrix that is nearly uncoupled with respect to $\epsilon$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Let $\mathcal{E}$ be an almost invariant aggregate of order $m$ contained in the associated state space and suppose that $\pi_i$ is constant for $i \in \mathcal{E}$. Let $\delta < 1$ satisfy*

$$\delta \geq \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon}$$

*and let the digraph $G$ be formed by an application of the maximum entry algorithm with inputs $A$ and $\delta$. Then, the algorithm will not remove every member of $\mathcal{E}$. Moreover, for any $i \in \mathcal{E}$ that is removed by the algorithm, the directed arc $i \to j$ present in $G$ has $j \in \mathcal{E}$. That is, any states in $\mathcal{E}$ that are selected for removal by the algorithm will be correctly associated with other members of $\mathcal{E}$.*

## 5.6 Continuity conditions concerning the maximum entry algorithm

We present two results concerning the robustness of the maximum entry algorithm. The results in this section are also true of the other stochastic complement based algorithms we present in later sections; only slight modifications to the proofs are required to show that the statements herein are true of all of our proposed algorithms. This is in contrast to the results in the previous section – proving that Proposition 5.10, or some similar statement, holds for our other algorithms does not

seem possible without extensive further assumptions on the matrix involved.

### 5.6.1 Uncoupled stochastic matrices and the maximum entry algorithm

First, we show that if the maximum entry algorithm is run on a stochastic matrix $A$ with input value $\delta = 0$, then there is a one-to-one correspondence between the weakly connected components of the output digraph and the essential classes of states of the associated Markov chain.

**Proposition 5.12.** *Let $A$ be a stochastic matrix on the state space $\mathcal{S}$. Let $\mathcal{E}_1, \ldots, \mathcal{E}_m$ be the essential classes of states contained in the associated state space and let $\mathcal{E}_0$ be the the collection of transient states. Let the digraph $G$ be formed by an application of Algorithm 3 with inputs $A$ and $\delta = 0$. Let $\mathcal{E} \subseteq \mathcal{S}$ be a weakly connected component of the digraph $G$; then, there is a unique $k \neq 0$ such that*

$$\mathcal{E}_k \subseteq \mathcal{E} \subseteq \mathcal{E}_k \cup \mathcal{E}_0.$$

**Remark.** Let $A$, $G$, $\mathcal{S}$ and $\Psi = \{\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0\}$ be as above. The vertex sets of the weakly connected components of $G$ form a partition of $\mathcal{S}$. The above proposition informs us that there is a partition of $\mathcal{E}_0$ into $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$, where the members $\mathcal{C}_k$ may be empty, such that the weakly connected components of $G$ are the induced subgraphs on the collections $\mathcal{E}_k \cup \mathcal{C}_k$.

**Proof** In light of Proposition 5.4, we simply need to show that

1. for every directed arc $i \to j$ present in $G$, either $i$ is transient or $i$ and $j$ are contained in the same essential class of states, and

2. any two states with out-degree equal to 0 in $G$ are contained in distinct essential classes of states.

The second condition implicitly implies that any vertex with out-degree equal to 0 in $G$ is not a member of $\mathcal{E}_0$. When it holds, the unique member of a weakly connected component $\mathcal{E}$ with out-degree 0 will not be contained in $\mathcal{E}_0$ and thus the vertex set of a weakly connected component is not a subset of $\mathcal{E}_0$.

Let $X$ be the Markov chain on $\mathcal{S}$ with transition matrix $A$. Recall that for $i, j \in \mathcal{S}$, we use the notation $i \prec j$ to represent that it is possible for the Markov chain to visit first $i$ and then, after 1 one or more transitions, $j$.

First suppose that the arc $i \to j$ is present in $G$. Then, the algorithm constructed, after some iteration, a stochastic complement of $A$ with the $ij$th entry not equal to 0. By Proposition 4.3, we must have $i \prec j$. If state $i$ is recurrent, then $i$ and $j$ are contained in the same essential class of states (see the proof of Theorem 1.7). So, either $i$ is transient (not recurrent) or $i$ and $j$ are contained in the same essential class of states.

Now, let $j_1, \ldots, j_{m'}$ be the members of $\mathcal{S}$ that have out-degree equal to 0. During its final iteration, the algorithm constructed the stochastic complement

$$B = S(\{j_1, \ldots, j_{m'}\})$$

and then terminated. Either $B = [1]$ or every off-diagonal entry of $B$ is less than or equal to the input value $\delta$. Since we are considering $\delta = 0$, in either case

$$S(\{j_1, \ldots, j_{m'}\}) = I.$$

By Proposition 4.11, the collection $\{j_1, \ldots, j_{m'}\}$ consists of one member from each essential class. ∎

This result is, in a sense, the motivation for utilising the stochastic complement in such a manner. Let $A$ be a stochastic matrix on the state space $\mathcal{S}$. An almost invariant aggregate is a collection of states that is nearly essential. When the algorithm is run with input $\delta = 0$, it reduces $A$, via successive stochastic complements, to the identity and produces a digraph on $G$ where each directed arc represents a transition within an essential class or beginning with a transient. When run with input $\delta > 0$, but still sufficiently small, the algorithm reduces $A$ to a stochastic complement that is near the identity. We then look to the digraph to construct candidate membership classes for almost invariant aggregates.

### 5.6.2 A continuity result concerning the maximum entry algorithm

We show that if a Markov chain is sufficiently uncoupled, the Maximum Entry Algorithm will produce accurate results.

We will make use of the $\infty$-norm on a matrix: given $A \in \mathbb{C}^{m\times n}$,

$$\|A\|_\infty = \max_{1\le i\le m} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}.$$

The $\infty$-norm is often referred to as the *maximum absolute row sum.* We note that for all $i$ and $j$,

$$|a_{ij}| \le \sum_k |a_{ik}| \le \|A\|_\infty .$$

So, for any two matrices $A$ and $B$ of the same order, $|b_{ij} - a_{ij}| \le \|B - A\|_\infty$, for all $i$ and $j$.

**Lemma 5.13.** *Let $A$ be a stochastic matrix and let $A \setminus \mathcal{C}$ be a stochastic complement of $A$. Then, there is an open neighbourhood of $A$ over which the map*

$$\tilde{A} \mapsto \tilde{A} \setminus \mathcal{C}$$

*is defined and continuous.*

**Proof** If $A = I$, then we must have $\mathcal{C} = \emptyset$; in this case, the statement is trivial as the described map is the identity.

Assume that $A \neq I$ and let $\beta$ be the smallest positive value among the off-diagonal entries of $A$. We show that the map

$$\tilde{A} \mapsto \tilde{A} \setminus \mathcal{C}$$

is continuous on the open set

$$\mathcal{B} = \left\{ \tilde{A} : \left\| \tilde{A} - A \right\|_\infty < \beta \right\}.$$

Let $\tilde{A} \in \mathcal{B}$. If $a_{ij} > 0$ and $i \neq j$, then we have $a_{ij} \geq \beta$. So, for all $i \neq j$ such that $a_{ij} \neq 0$,

$$|\tilde{a}_{ij} - a_{ij}| < \beta \leq a_{ij},$$

in which case $\tilde{a}_{ij} > 0$. In other words, in any off-diagonal position where $A$ is nonzero, $\tilde{A}$ is nonzero as well. Now, let $i \in \mathcal{C}$. Since $A \setminus \mathcal{C}$ exists, $\mathcal{C}$ does not contain an entire essential class of states, with respect to $A$. By Theorem 1.8, there must be a state $j \notin \mathcal{C}$ and a sequence

$$i = i_0, i_1, \ldots, i_l = j$$

such that $i_s \neq i_{s+1}$ and $a_{i_s i_{s+1}} > 0$ for $s = 0, \ldots, l-1$. Thus, $\tilde{a}_{i_s i_{s+1}} > 0$ for $s = 0, \ldots, l-1$. So, $\mathcal{C}$ does not contain an entire essential class of states with respect to $\tilde{A}$, either. Thus, the map

$$\tilde{A} \mapsto \tilde{A} \setminus \mathcal{C}$$

is defined on the open set $\mathcal{B}$. The entries of $\tilde{A} \setminus \mathcal{C}$ are rational functions of the entries of $\tilde{A}$. A rational function is continuous on any set over which it is defined. Thus, the given map is entrywise continuous over $\mathcal{B}$. A (finite-dimensional) matrix function which is entrywise continuous is continuous under the $\infty$-norm. ∎

Recall that if $A$ is a reversible substochastic matrix, then there is a positive diagonal matrix $\Pi$ such that $\Pi A$ is symmetric. Moreover, if $A$ is irreducible, such a matrix is uniquely determined, up to multiplication by a positive constant.

**Lemma 5.14.** *Let $A$ be an irreducible reversible stochastic matrix and let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Then, for any positive value $\epsilon$, there is a positive value $\delta$ such that if*

1. *$\tilde{A}$ is a reversible substochastic matrix of the same order as $A$,*
2. *$\tilde{\Pi}$ is a positive diagonal matrix such that $\tilde{\Pi}\tilde{A}$ is symmetric, and*
3. *$\|\tilde{A} - A\|_\infty < \delta$,*

*then for all $i$ and $j$,*

$$\left| \frac{\tilde{\pi}_i}{\tilde{\pi}_j} - \frac{\pi_i}{\pi_j} \right| < \epsilon.$$

**Proof** If $A$ has order 1, then the statement is trivial, as the value

$$\left| \frac{\tilde{\pi}_i}{\tilde{\pi}_j} - \frac{\pi_i}{\pi_j} \right|$$

is nonzero only in the case that $i \neq j$. So assume that $A$ is irreducible and of order 2 or greater; let $\beta$ be the smallest positive value among the off-diagonal entries of $A$. Suppose that

1. $\tilde{A}$ is a reversible substochastic matrix of the same order as $A$,

2. $\tilde{\Pi}$ is a positive diagonal matrix such that $\tilde{\Pi}\tilde{A}$ is symmetric, and

3. $\|\tilde{A} - A\|_\infty < \beta$.

Since $\tilde{\Pi}\tilde{A}$ is symmetric, whenever $i \neq j$ and $\tilde{a}_{ij} > 0$, we have $\tilde{a}_{ji} > 0$ and

$$\frac{\tilde{\pi}_i}{\tilde{\pi}_j} = \frac{\tilde{a}_{ji}}{\tilde{a}_{ij}}.$$

As in the proof of Lemma 5.13, whenever $a_{ij} > 0$, we have $\tilde{a}_{ij} > 0$; so, the matrix $\tilde{A}$ is irreducible. Let $i \neq j$; then, there is a sequence

$$i = i_0, i_1, \ldots, i_l = j$$

such that $i_s \neq i_{s+1}$ and $\tilde{a}_{i_s i_{s+1}} > 0$ for $s = 0, \ldots, l-1$. Thus,

$$\frac{\tilde{\pi}_i}{\tilde{\pi}_j} = \frac{\tilde{\pi}_{i_0}\tilde{\pi}_{i_1}\cdots\tilde{\pi}_{i_{l-1}}}{\tilde{\pi}_{i_1}\tilde{\pi}_{i_2}\cdots\tilde{\pi}_{i_l}} = \frac{\tilde{a}_{i_1 i_0}\tilde{a}_{i_2 i_1}\cdots\tilde{a}_{i_l i_{l-1}}}{\tilde{a}_{i_0 i_1}\tilde{a}_{i_1 i_2}\cdots\tilde{a}_{i_{l-1} i_l}}.$$

So, the ratios $\tilde{\pi}_i/\tilde{\pi}_j$ are continuous functions of the entries of $\tilde{A}$. There are only finitely many such ratios, so for any $\epsilon > 0$, there is $\delta \leq \beta$ such that if $\tilde{A}$ is reversible and $\|\tilde{A} - A\|_\infty < \delta$ then for all $i$ and $j$,

$$\left| \frac{\tilde{\pi}_i}{\tilde{\pi}_j} - \frac{\pi_i}{\pi_j} \right| < \epsilon.$$

∎

**Proposition 5.15.** *Let*

$$A \cong \begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_m \end{bmatrix},$$

*where $m \geq 2$ and each $B_k$ is an irreducible reversible stochastic matrix. Let $\mathcal{S}$ be the associated state space and for each $k$ let $\mathcal{E}_k \subseteq \mathcal{S}$ be the collection of states associated with block $B_k$.*

*There are positive values $\delta$ and $d$ such that for any reversible stochastic matrix $\tilde{A}$ on $\mathcal{S}$ with $\|\tilde{A} - A\|_\infty < \delta$, the Maximum Entry Algorithm, with inputs $\tilde{A}$ and $d$, will return a digraph whose weakly connected components are the $m$ induced subgraphs on the collections $\mathcal{E}_k$.*

**Proof** If $A = I$, the claim is true simply by selecting $\delta = d$. Suppose that $\|\tilde{A} - I\|_\infty < \delta$ and that the digraph $G$ is formed by an application of the Maximum Entry Algorithm to $\tilde{A}$ with input value $\delta$. Since $\|\tilde{A} - I\|_\infty < \delta$, we have $\tilde{a}_{ij} < \delta$ for all $i \neq j$. Thus, the algorithm terminates without adding a single arc to $G$.

So, suppose that $A \neq I$. Let $\Sigma$ be the collection of subsets $\mathcal{C} \subseteq \mathcal{S}$ such that

1. the stochastic complement $A \setminus \mathcal{C}$ exists and

2. $A \setminus \mathcal{C} \neq I$.

By Propositions 4.2 and 4.11, $\mathcal{C} \in \Sigma$ if and only if

1. for all $k$, $\mathcal{E}_k \not\subseteq \mathcal{C}$, and

2. for at least one $k$, $|\mathcal{E}_k \setminus \mathcal{C}| \geq 2$.

We note that $\emptyset \in \Sigma$.

Let

$$\beta = \min_{\mathcal{C} \in \Sigma} \max_{\substack{i,j \notin \mathcal{C} \\ \exists j \neq i}} \left\{ (A \setminus \mathcal{C})_{ij} \right\}.$$

Since $A \setminus \mathcal{C} \neq I$ for all $\mathcal{C} \in \Sigma$, $\beta > 0$. For all $\mathcal{C} \in \Sigma$, $A \setminus \mathcal{C}$ has at least one off-diagonal entry greater than or equal to $\beta$.

For each $\mathcal{C} \in \Sigma$, let $\delta_{\mathcal{C}}$ be such that if $\tilde{A}$ is stochastic and $\|\tilde{A} - A\|_\infty < \delta_{\mathcal{C}}$, then $\tilde{A} \setminus \mathcal{C}$ exists and $\|\tilde{A} \setminus \mathcal{C} - A \setminus \mathcal{C}\|_\infty < \beta/2$. Let $\delta' = \min_{\mathcal{C} \in \Sigma}\{\delta_{\mathcal{C}}\}$. So, if $\|\tilde{A} - A\|_\infty < \delta'$, then for all $\mathcal{C} \in \Sigma$, $\tilde{A} \setminus \mathcal{C}$ exists and $\|\tilde{A} \setminus \mathcal{C} - A \setminus \mathcal{C}\|_\infty < \beta/2$. We note that this implies that if $\|\tilde{A} - A\|_\infty < \delta'$, then for all $\mathcal{C} \in \Sigma$, the largest off-diagonal entry of $\tilde{A} \setminus \mathcal{C}$ is greater than or equal to $\beta/2$.

For $k = 1, \ldots, m$, let $\Pi^{(k)}$ be a positive diagonal matrix such that $\Pi^{(k)} B_k$ is symmetric. Let

$$\rho = \max_{\substack{i,j,k \\ \exists i,j \in \mathcal{E}_k}} \left\{ \frac{\pi_i^{(k)}}{\pi_j^{(k)}} \right\}.$$

We note that $\rho \geq 1$. For each $k$, let $\delta_k$ be such that if

1. $\tilde{B}$ is a reversible substochastic matrix of order equal to $B_k$,

2. $\tilde{\Pi}$ is a positive diagonal matrix such that $\tilde{\Pi}\tilde{B}$ is symmetric and

3. $\|\tilde{B} - B_k\|_\infty < \delta_k$,

then, for all $i, j \in \mathcal{E}_k$,

$$\left| \frac{\tilde{\pi}_i}{\tilde{\pi}_j} - \frac{\pi_i^{(k)}}{\pi_j^{(k)}} \right| < \rho.$$

Note that if $\tilde{\Pi}$ is as above, the above inequality implies that for any $i$ and $j$,

$$\frac{\tilde{\pi}_{ii}}{\tilde{\pi}_{jj}} < \rho + \frac{\pi_{ii}^{(k)}}{\pi_{jj}^{(k)}} \leq 2\rho.$$

Let $\delta'' = \min_{1 \leq k \leq m}\{\delta_k\}$.

Finally, let

$$\delta''' = \min_{1 \leq k \leq m} \left\{ \frac{\beta}{4p\,|\mathcal{E}_k|} \right\}$$

and let $\delta = \min\{\delta', \delta'', \delta'''\}$.

Let $\tilde{A}$ be a reversible stochastic matrix such that $\|\tilde{A} - A\|_\infty < \delta$, and let the digraph $G$ be formed by an application of the Maximum Entry Algorithm to $\tilde{A}$ with input value $d = \beta/2$. We claim that vertex sets of the weakly connected components of $G$ are the collections $\{\mathcal{E}_k\}$.

We will first show that if the directed arc $i \to j$ is present in $G$, then the states $i$ and $j$ are members of the same aggregate $\mathcal{E}_k$. We will accomplish this by showing that for each $\mathcal{E}_k$, if $\mathcal{C} \subseteq \mathcal{S}$ is such that

1. $\tilde{A} \setminus \mathcal{C}$ exists and

2. $\mathcal{E}_k \setminus \mathcal{C}$ is nonempty,

then

$$\sum_{i \in \mathcal{E}_k} \sum_{j \notin \mathcal{E}_k \cup \mathcal{C}} \left( \tilde{A} \setminus \mathcal{C} \right)_{ij} < \frac{\beta}{2}.$$

This is sufficient, because the directed arc $i \to j$ can be present in $G$ only if there is a stochastic complement $\tilde{A} \setminus \mathcal{C}$ such that $i, j \notin \mathcal{C}$ and

$$\left( \tilde{A} \setminus \mathcal{C} \right)_{ij} > \frac{\beta}{2}.$$

Let $\mathcal{E}_k$ be one of the aggregates of $A$. We note that since $a_{ij} = 0$ for all $i \in \mathcal{E}_k$ and $j \notin \mathcal{E}_k$ and $\|\tilde{A} - A\|_\infty < \delta \leq \delta'''$, for every $i \in \mathcal{E}_k$,

$$\sum_{j \notin \mathcal{E}_k} \tilde{a}_{ij} = \sum_{j \notin \mathcal{E}_k} |\tilde{a}_{ij} - a_{ij}| < \delta''' \leq \frac{\beta}{4p \left| \mathcal{E}_k \right|}.$$

First, suppose that $\mathcal{C} \cap \mathcal{E}_k = \emptyset$. Then, the principle submatrix of $\tilde{A} \setminus \mathcal{C}$ corresponding to $\mathcal{E}_k$ is bounded below, entrywise, by the principal submatrix of $\tilde{A}$ corresponding to $\mathcal{E}_k$. Thus,

$$\begin{aligned}
\sum_{i\in\mathcal{E}_k}\sum_{j\notin\mathcal{E}_k\cup\mathcal{C}}\left(\tilde{A}\setminus\mathcal{C}\right)_{ij} &= |\mathcal{E}_k| - \sum_{i\in\mathcal{E}_k}\sum_{j\in\mathcal{E}_k}\left(\tilde{A}\setminus\mathcal{C}\right)_{ij} \\
&\leq |\mathcal{E}_k| - \sum_{i\in\mathcal{E}_k}\sum_{j\in\mathcal{E}_k}\tilde{a}_{ij} \\
&= \sum_{i\in\mathcal{E}_k}\sum_{j\notin\mathcal{E}_k}\tilde{a}_{ij} \\
&\leq |\mathcal{E}_k|\delta''' \\
&\leq |\mathcal{E}_k|\beta/(4p|\mathcal{E}_k|) \\
&< \beta/2
\end{aligned}$$

$(2 < 4p)$. Now, suppose that $\mathcal{C} \subseteq \mathcal{E}_k$. Express

$$\tilde{A} \cong \begin{bmatrix} A_{11} & A_{12} & F_1 \\ A_{21} & A_{22} & F_2 \\ * & * & * \end{bmatrix},$$

where the first position corresponds to $\mathcal{E}_k \setminus \mathcal{C}$, the second to $\mathcal{C}$, and the third to $\mathcal{S} \setminus \mathcal{E}_k$ (only the first two rows of blocks will appear in our calculations). As we noted above, for any $i \in \mathcal{E}_k$,

$$\sum_{j\notin\mathcal{E}_k} a_{ij} < \delta.$$

So, $F_1\mathbb{1} \leq \delta\mathbb{1}$ and $F_2\mathbb{1} \leq \delta\mathbb{1}$. Now,

$$\tilde{A}\setminus\mathcal{C} \cong \begin{bmatrix} A_{11} + A_{12}(I - A_{22})^{-1}A_{21} & F_1 + A_{12}(I - A_{22})^{-1}F_2 \\ * & * \end{bmatrix}.$$

Let $\tilde{D}$ be a positive diagonal matrix such that $\tilde{D}\tilde{A}$ is symmetric; let $\tilde{D}_1$ and $\tilde{D}_2$ be the principal submatrices corresponding to $\mathcal{E}_k \setminus \mathcal{C}$ and $\mathcal{C}$, respectively. So, $\tilde{D}_2 A_{22} = A_{22}^T \tilde{D}_2$ and $\tilde{D}_1 A_{12} = A_{21}^T \tilde{D}_2$.

We note that

$$\left\| \tilde{A}(\mathcal{E}_k, \mathcal{E}_k) - B_k \right\|_\infty \leq \left\| \tilde{A} - A \right\|_\infty < \delta'';$$

so,

$$\frac{\tilde{d}_i}{\tilde{d}_j} < 2p,$$

for all $i, j \in \mathcal{E}_k$. As well, if $Y$ as a nonnegative matrix of the appropriate order,

$$\tilde{D}_1^{-1} Y \tilde{D}_2 = \left[ \frac{\left(\tilde{D}_2\right)_{jj}}{\left(\tilde{D}_1\right)_{ii}} y_{ij} \right] < [2py_{ij}] = 2pY.$$

Since $\tilde{A}$ is stochastic, $A_{21}\mathbb{1} + A_{22}\mathbb{1} \leq \mathbb{1}$, further implying that $(I - A_{22})^{-1} A_{21}\mathbb{1} \leq \mathbb{1}$. We rewrite this inequality as

$$\mathbb{1}^T A_{21}^T \left(I - A_{22}^T\right)^{-1} \leq \mathbb{1}^T.$$

Thus,

$$\begin{aligned}
\sum_{i\in\mathcal{E}_k\setminus\mathcal{C}}\sum_{j\notin\mathcal{E}_k\setminus\mathcal{C}}\left(\tilde{A}\setminus\mathcal{C}\right)_{ij} &= \mathbb{1}^T\left(F_1+A_{12}(I-A_{22})^{-1}F_2\right)\mathbb{1}\\
&= \mathbb{1}^T\left(F_1+\tilde{D}_1^{-1}A_{21}^T\tilde{D}_2(I-A_{22})^{-1}F_2\right)\mathbb{1}\\
&= \mathbb{1}^T\left(F_1+\tilde{D}_1^{-1}A_{21}^T(I-A_{22}^T)^{-1}\tilde{D}_2F_2\right)\mathbb{1}\\
&< \mathbb{1}^T\left(F_1+2pA_{21}^T(I-A_{22}^T)^{-1}F_2\right)\mathbb{1}\\
&= \mathbb{1}^TF_1\mathbb{1}+2p\mathbb{1}^TA_{21}^T(I-A_{22}^T)^{-1}F_2\mathbb{1}\\
&\leq \mathbb{1}^TF_1\mathbb{1}+2p\mathbb{1}^TF_2\mathbb{1}\\
&< \delta''\left|\mathcal{E}_k\setminus\mathcal{C}\right|+2p\delta''\left|\mathcal{C}\right|\\
&< 2p\delta''\left|\mathcal{E}_k\setminus\mathcal{C}\right|+2p\delta''\left|\mathcal{C}\right|\\
&= 2p\delta''\left|\mathcal{E}_k\right|.
\end{aligned}$$

(The second to last inequality is arrived at by noting that $p\geq 1$). Now,

$$\delta''\leq\frac{\beta}{4p\left|\mathcal{E}_k\right|},$$

implying that

$$\begin{aligned}
\sum_{i\in\mathcal{E}_k\setminus\mathcal{C}}\sum_{j\notin\mathcal{E}_k\setminus\mathcal{C}}\left(\tilde{A}\setminus\mathcal{C}\right)_{ij} &< 2p\delta''\left|\mathcal{E}_k\right|\\
&\leq \tfrac{\beta}{2}.
\end{aligned}$$

Finally, suppose that $\mathcal{C}'=\mathcal{C}\cap\mathcal{E}_k$ and $\mathcal{C}''=\mathcal{C}\setminus\mathcal{E}_k$ are both nonempty. By our above reasoning,

$$\sum_{i\in\mathcal{E}_k\setminus\mathcal{C}'}\sum_{j\notin\mathcal{E}_k}\left(\tilde{A}\setminus\mathcal{C}'\right)_{ij}<\frac{\beta}{2}.$$

Then, using the fact that the principal submatrix of $\tilde{A} \setminus \mathcal{C}$ corresponding to $\mathcal{E}_k \setminus \mathcal{C}$ is bounded below, entrywise, by the principal submatrix of $\tilde{A} \setminus \mathcal{C}'$ corresponding to $\mathcal{E}_k \setminus \mathcal{C}$, we see that

$$\begin{aligned}
\sum_{i \in \mathcal{E}_k \setminus \mathcal{C}} \sum_{j \notin \mathcal{E}_k \cup \mathcal{C}} \left( \tilde{A} \setminus \mathcal{C} \right)_{ij} &= |\mathcal{E}_k \setminus \mathcal{C}| - \sum_{i \in \mathcal{E}_k \setminus \mathcal{C}} \sum_{j \in \mathcal{E}_k \setminus \mathcal{C}} \left( \tilde{A} \setminus \mathcal{C} \right)_{ij} \\
&\leq |\mathcal{E}_k \setminus \mathcal{C}'| - \sum_{i \in \mathcal{E}_k \setminus \mathcal{C}'} \sum_{j \in \mathcal{E}_k \setminus \mathcal{C}'} \left( \tilde{A} \setminus \mathcal{C}' \right)_{ij} \\
&= \sum_{i \in \mathcal{E}_k \setminus \mathcal{C}'} \sum_{j \notin \mathcal{E}_k} \left( \tilde{A} \setminus \mathcal{C}' \right)_{ij} \\
&< \tfrac{\beta}{2}.
\end{aligned}$$

Therefore, if the directed arc $i \to j$ is present in the output digraph $G$, then the states $i$ and $j$ are contained in the same aggregate $\mathcal{E}_k$.

Suppose that the Maximum Entry Algorithm, applied to $\tilde{A}$, has executed $s$ iterations and let $G_s$ be the digraph at this point. By Lemma 5.3, the number of weakly connected components of $G_s$ is $n - s$, where $n$ is the order of $\tilde{G}$. By our above reasoning, each of the weakly connected components of $G$ contains states from exactly one of the collections $\mathcal{E}_1, \ldots, \mathcal{E}_m$. Thus, it is now sufficient to show that the algorithm executes at least $s = n - m$ iterations.

Suppose that the algorithm has executed $s < n - m$ iterations; let $\mathcal{C}$ be the collection of states removed, via stochastic complements, so far and let $G_s$ be the current digraph. The digraph $G_s$ is acyclic and $i \prec_{G_s} j$ implies that $i, j \in \mathcal{E}_k$ for some $k$. So, for each $k$, there is $j \in \mathcal{E}_k$ such that $j \notin \mathcal{C}$. Thus, for all $k$, $\mathcal{E}_k \nsubseteq \mathcal{C}$. As well,

$$s = |\mathcal{C}| = \sum_{k=1}^{m} |\mathcal{C} \cap \mathcal{E}_k| < n - m = \sum_{k=1}^{m} (|\mathcal{E}_k| - 1) .$$

So, for at least one $k$, $|\mathcal{C} \cap \mathcal{E}_k| < |\mathcal{E}_k| - 1$, in which case $|\mathcal{E}_k \setminus \mathcal{C}| \geq 2$. So, $\mathcal{C} \in \Sigma$ (described above). Since $\|\tilde{A} - A\|_\infty < \delta'$, there is an off-diagonal entry of $\tilde{A} \setminus \mathcal{C}$ strictly greater than $\beta/2$. Thus, after, with input $d = \beta/2$, after executing $s < n - m$ iterations, the algorithm executes at least one more iteration.

Therefore, if $\|\tilde{A} - A\|_\infty < \delta$ and the digraph $G$ is obtained by an application of the Maximum Entry Algorithm with input $d = \beta/2$, then each weakly connected component of $G$ contains states from exactly one of the collections $\mathcal{E}_1, \ldots, \mathcal{E}_m$ and the digraph $G$ contains exactly $m$ weakly connected components. The vertex sets of the weakly connected components of $G$ must be the aggregates $\mathcal{E}_k$. ∎

## 5.7 The modified maximum entry algorithm

Let $A$ be a reversible stochastic matrix with state space $\mathcal{S}$ and let $\epsilon$ be a positive number strictly less than 1. If we want to test whether $A$ is nearly uncoupled with respect to $\epsilon$ using the maximum entry algorithm, we need to select an appropriate input value for $\delta$. If one knows, *a priori,* the sizes of the almost invariant aggregates, or at least has an approximate lower bound for their sizes, Proposition 5.10 can be utilised to select an appropriate $\delta$. Moreover, if an arbitrary lower bound is set, this can be used.

**Algorithm 4** The modified maximum entry algorithm

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that, initially, contains no arcs.
$\mathcal{C} := \emptyset$
$m := \mathbb{1}_{\mathcal{S}}$
**while** the order of $B$ is 2 or greater **do**
  Let $i, j \in \mathcal{S} \setminus \mathcal{C}$ be such that $i \neq j$ and $b_{ij} = \max_{j' \neq i'} \{b_{i'j'}\}$.
  **if** $b_{ii} \geq \frac{(1-\epsilon)^2}{1+(m_i-2)\epsilon}$ **then**
    Exit the **while** loop.
  **else**
    Add the directed arc $i \to j$ to $G$.
    $B := B \setminus i$
    $m_j := m_j + m_i$
    $\mathcal{C} := \mathcal{C} \cup \{i\}$
  **end if**
**end while**
**return** $G$

For example, if one wishes to construct candidate subsets $\mathcal{E} \subseteq \mathcal{S}$ that have size at least $m$ and are almost invariant aggregates with respect to $\epsilon$, then the input value

$$\delta = \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon}$$

(or, more conservatively, $\delta = \epsilon m$) may be used.

However, if no such lower bound on the order is known (or desirable) it is difficult to select an appropriate $\delta$. Thus, we present a modified version of the maximum entry algorithm.

Rather than selecting a maximal entry $b_{ij}$ and then comparing it with $\delta$ to decide whether this entry represents a transition within an almost invariant aggregate, this algorithm utilises a test to determine whether the state $i$ is "safe" to remove. This

version takes as inputs a reversible stochastic matrix $A$ on the state space $\mathcal{S}$ and a value $\epsilon < 1$ and attempts to construct candidate subsets of the state space which are almost invariant with respect to $\epsilon$.

The vector $\mathbb{1}_{\mathcal{S}}$ utilised within Algorithm 4 is the vector indexed by $\mathcal{S}$ that has every entry equal to 1.

As with Algorithm 3, we refer to execution of the four commands after the **else** statement as an iteration of the algorithm.

**Proposition 5.16.** *Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $0 \leq \epsilon < 1$; suppose that we have applied Algorithm 4 with inputs $A$ and $\epsilon$. Let $r$ be the number of iterations completed by the algorithm and let $B$, $\mathcal{C}$, $G$ and $m$ be the stored data after $s \leq r$ iterations. Then,*

1. *$\mathcal{C}$ contains $s$ states and $B \cong A \setminus \mathcal{C}$,*
2. *$G$ is acyclic and contains $s$ directed arcs,*
3. *every member of $\mathcal{C}$ has out-degree 1 in $G$ and every member of $\mathcal{S} \setminus \mathcal{C}$ has out-degree 0, and*
4. *for each $i \in \mathcal{S} \setminus \mathcal{C}$, $m(i)$ is the number of states contained in the weakly connected component of $G$ which contains $i$.*

**Proof** The first three statements are shown in the same manner as in Proposition 5.1 and Lemma 5.3. We prove the fourth by induction on $s$.

For $s = 0$, $G$ contains no arcs and so every weakly connected component of $G$ consists of a single isolated vertex. We have $m_i = 1$, for all $i \in \mathcal{S}$, so the statement holds.

Now, let $1 \leq s \leq r$ and let $\mathcal{C}'$, $\mathcal{C}$, $m'$, $m$, $G'$ and $G$ be the stored data after $s' = s - 1$ and $s$ iterations, respectively. Let

$$\mathcal{S} \setminus \mathcal{C}' = \{j_1, \ldots, j_n\}$$

be the states with out-degree 0 in $G'$. Thus, each $j_k$ is contained in a distinct weakly connected component of $G'$, say $\mathcal{E}'_k$. So, for $k = 1, \ldots, n$, $m'(j_k) = |\mathcal{E}'_k|$.

During the $s$th iteration, the algorithm selects distinct states $j_k$ and $j_l$ with out-degree 0 and forms $\mathcal{C}$, $m$ and $G$ by adding the arc $j_k \to j_l$ to $G'$, adding $j_k$ to $\mathcal{C}'$ and replacing $m'_{j_l}$ with $m'_{j_k} + m'_{j_l}$.

Without loss of generality, we assume that the directed arc $j_n \to j_{n-1}$ is added to $G'$. This increases the degree of $j_n$ from 0 to 1 and merges the weakly connected components $\mathcal{E}'_n$ and $\mathcal{E}'_{n-1}$. So, the weakly connected components of $G$ are $\mathcal{E}_1, \ldots, \mathcal{E}_{n-1}$ where $\mathcal{E}_k = \mathcal{E}'_k$ if $k \leq n - 2$ and $\mathcal{E}_{n-1} = \mathcal{E}'_{n-1} \cup \mathcal{E}'_n$. The unique member of $\mathcal{E}_k$ with out-degree 0 is $j_k$. So, for $k = 1, \ldots, n - 1$,

$$m_{j_k} = \begin{cases} m'_{j_k} & \text{if } k \neq n-1 \\ m'_{j_{n-1}} + m'_{j_n} & \text{if } k = n-1. \end{cases}$$

Thus, for all $i$ with out-degree equal to 0 in $G$, $m_i$ is the number of states contained in the same weakly connected component as $i$. ∎

Let $G$ be a reversible stochastic matrix with state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate, with respect to $\epsilon > 0$, containing $m$ states. Suppose that the maximum entry algorithm has been applied to $A$ and that, after some number of iterations, $m-1$ states contained in $\mathcal{E}$ have been removed via stochastic complements.

Let $i \in \mathcal{E}$ be the state that has not yet been removed and suppose further that all the members of $\mathcal{E}$ removed so far have been correctly associated with other members of $A$. That is, suppose that for all $i' \in \mathcal{E} \setminus i$, the directed arc $i' \to j'$ present in the constructed digraph has $j' \in \mathcal{E}$. Then, at this point, we must have $m_i \geq m$ (some near-transient states may have been associated with members of $\mathcal{E}$). So, using Lemma 5.9 and Proposition A.14, as in the proof of Proposition 5.10, we have

$$b_{ii} \geq \frac{(1-\epsilon)^2}{1+(m-2)\epsilon} \geq \frac{(1-\epsilon)^2}{1+(m_i-2)\epsilon},$$

where $B$ is the stochastic complement of $A$ currently under consideration.

The reasoning behind the steps of the modified maximum entry algorithm is the following. At a given iteration, let $A$ be the stochastic complement currently under consideration.

1. The algorithm identifies the maximal off-diagonal entry $b_{ij}$ – this pair of states is the most likely to be part of the same almost invariant aggregate of the original Markov chain.

2. If

$$b_{ii} < \frac{(1-\epsilon)^2}{1+(m(i)-2)\epsilon},$$

we suspect that either $i$ is near transient or there are further states, not yet removed, contained in the same almost invariant aggregate as $i$. (The above inequality suggests that $i$ is not the final unremoved member of an almost invariant aggregate). In either case, it is safe to remove $i$ without disrupting the uncoupled structure of the matrix; we reason that if $i$ is not near transient, the most likely candidate for a state contained in the same aggregate as $i$ is $j$ (identified above).

3. If

$$b_{ii} \geq \frac{(1-\epsilon)^2}{1+(m(i)-2)\epsilon},$$

the state $i$ may be the final, not yet removed, member of an almost invariant aggregate. We cannot be confident that removing $i$ will not disrupt the structure of the matrix. Moreover, at this point we have identified the very largest off-diagonal entry in the matrix $A$ and discovered that it may represent a transition between members of distinct aggregates. If the very largest entry is such, we reason that all of the entries may represent transitions between aggregates; so, we terminate the algorithm in this occurrence.

We emphasise that this line of reasoning relies on the assumption that the largest entry in the reversible matrix under consideration does identify a transition between members of the same aggregate (or from a near transient state to another state). This is not always the case; for instance, see Example 5.17, below. However, in worked examples, it seems that every correct association the algorithm adds to the digraph increases the likelihood of further correct associations.

**Example 5.17.** Let $\epsilon$ be a positive constant very near to 0 and let $m \geq 1/\epsilon$ be a positive integer. Consider the reversible stochastic matrix

$$A = \begin{bmatrix} \frac{1-\epsilon}{m} J & \epsilon \mathbb{1} \\ \\ \frac{\epsilon}{m} \mathbb{1}^T & 1-\epsilon \end{bmatrix},$$

where $J$ is the $m \times m$ matrix and $\mathbb{1}$ is the column vector of order $m$ with every entry equal to 1. Since $m \geq 1/\epsilon$, we have

$$m > \frac{1}{\epsilon} - 1 = \frac{1-\epsilon}{\epsilon},$$

further implying that $(1-\epsilon)/m < \epsilon$. The partition $(\{1, \ldots, m\}, \{m+1\})$ is the unique $\epsilon$-uncoupling of the state space of $A$. For any $i, j \in \{1, \ldots, m\}$, the transition $i \to m+1$ is more likely than the transition $i \to j$, even though $i$ and $j$ are contained in a minimal almost invariant aggregate and $i$ and $m+1$ are not.

## 5.8 Evaluating uncouplings of Markov chains

Let $A$ be a nearly uncoupled stochastic matrix with respect to $\epsilon > 0$ and let the digraph $G$ be produced by an application of the maximum entry algorithm. We present a method for determining whether the algorithm has been successful in its decoupling of the associated Markov chain.

Let $V$ be the vertex set of a weakly connected component of $G$. If the algorithm's output is correct, then $V$ consists of an almost invariant aggregate together with some collection of near transient states. Let $B = A(V)$ and let

$$\gamma = \gamma_B = (I - B)\mathbb{1}.$$

Simply calculating the value

$$\max_{i \in V} \{\gamma_i\}$$

is not a good indication of the algorithm's success or failure, as near transient members of $V$ are as likely to have large values in $\gamma$ as they are to have small values. We will instead use the $\mathbb{1}$-coupling measure

$$w_{\mathbb{1}}(B) = \frac{\mathbb{1}^T B \mathbb{1}}{|V|} = \frac{|V| - \mathbb{1}^T \gamma}{|V|}.$$

The value $w_{\mathbb{1}}(B)$ is the mean probability of transitioning from a member of $V$ to another member of $V$. That is,

$$w_{\mathbb{1}}(B) = \frac{1}{|V|} \sum_{i \in V} \mathbb{P}\left[x_{t+1} \in V : x_t = i\right].$$

Thus, when $\mathbb{1}$-coupling measures are each close to 1 (for all weakly connected components of the output digraph), we assume that the algorithm has performed well. This measure of the strength of the produced aggregates was introduced in [10].

In [18, 8, 7], the $\pi$-coupling measure is used to evaluate the strength of an aggregate. Let $A$ be a stochastic matrix with stationary distribution $\pi$. Let $V$ be a collection of states and let $B = A(V)$ and $u = \pi(V)$ be the principal submatrix and subvector, respectively, associated with $V$. The $\pi$-coupling measure is the value

$$w_\pi(B) = \frac{u^T B \mathbb{1}}{u^T \mathbb{1}} = \frac{u^T(\mathbb{1} - \gamma)}{u^T \mathbb{1}} = \frac{\sum\limits_{i \in V} \sum\limits_{j \in V} \pi_i a_{ij}}{\sum\limits_{i \in V} \pi_i}.$$

If the initial distribution is $\pi$, that is, if

$$\mathbb{P}\left[x_0 = i\right] = \pi_i$$

for all $i$ in the associated state space, then

$$w_\pi(B) = \mathbb{P}\left[x_{t+1} \in V | x_t \in V\right].$$

A somewhat straightforward application of Theorem 2.2 shows that if $A$ is irreducible

$$w_\pi(B) = \lim_{t \to \infty} \frac{1}{t+1} \sum_{s=0}^{t} \mathbb{P}\left[x_{t+1} \in V : x_t \in V\right],$$

regardless of the initial distribution. Thus, in principle, the $\pi$-coupling measure is a better indicator of whether or not a collection $V$ is decoupled from the remainder of the state space. However, the very fact that $A$ is nearly uncoupled implies that the vector $\pi$ is a difficult quantity to calculate (accurately). In general, we propose that the $\mathbb{1}$-coupling measure is a more practical indicator – it is fast and reliable to calculate. Moreover, it seems that for most matrices produced, $w_{\mathbb{1}}(B) \leq w_{\pi}(B)$ (although this is not necessarily the case); the $\mathbb{1}$-coupling measure seems to be, in practise, more conservative.

# Chapter 6

# Error-reducing algorithms

We present three algorithms which attempt to construct almost invariant aggregates of a given reversible stochastic matrix. These algorithms attempt to reduce or limit the growth of error terms (transitions between almost invariant aggregates) within the constructed stochastic complements.

## 6.1 Preliminaries

### 6.1.1 Error reduction in stochastic complements

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$ and let $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$ be an $\epsilon$-uncoupling of $A$. That is, for $k \neq 0$, $\mathcal{E}_k$ is a minimal almost invariant aggregate and, when it is nonempty, $\mathcal{E}_0$ does not contain any almost invariant aggregates as subsets.

We express $A \cong [A_{ij}]$ where $A_{ij} = A(\mathcal{E}_i, \mathcal{E}_j)$. For each $k$, let $B_k = A_{kk}$; for $k \neq 0$, the substochastic matrix $B_k$ is irreducible and

$$\gamma_{B_k} = (I - B_k)\mathbb{1} = \sum_{j \neq k} A_{kj}\mathbb{1} \leq \epsilon\mathbb{1}.$$

For $k \neq 0$, the total error at $\mathcal{E}_k$ is the number

$$\eta(B_k) = \mathbb{1}^T\gamma_{B_k} = \mathbb{1}^T(I - B_k)\mathbb{1}.$$

If $i \in \mathcal{E}_k$, where $k \neq 0$, $j \notin \mathcal{E}_k$ and $a_{ij} > 0$, we refer to $a_{ij}$ as an error term. Thus, the total error at $\mathcal{E}_k$ is the sum of the error terms contained in the rows corresponding to $\mathcal{E}_k$. Whenever the entry $a_{ij}$ is particularly large, we suspect that $a_{ij}$ is not an error term, in which case either $i, j \in \mathcal{E}_k$ for some $k \neq 0$, or $i \in \mathcal{E}_0$.

Let $\mathcal{C} \subseteq \mathcal{S}$ be such that for all $k \neq 0$, $\mathcal{E}_k \nsubseteq \mathcal{C}$. Then, $A \setminus \mathcal{C}$ exists, as any essential class of states must contain at least one of the collections $\mathcal{E}_k$ as a subset. Let $\hat{A} = A \setminus \mathcal{C}$ and for each $k \neq 0$, let $\hat{\mathcal{E}}_k = \mathcal{E}_k \setminus \mathcal{C}$ and $\hat{B}_k = \hat{A}(\hat{\mathcal{E}}_k)$.

We refer to the stochastic complement $\hat{A} = A \setminus \mathcal{C}$ as *error-reducing with respect to* $\Psi$ if, for all $k \neq 0$, $\eta(\hat{B}_k) \leq \eta(B_k)$. If $\hat{A}$ is an error-reducing stochastic complement and $\hat{a}_{ij}$ is relatively large, we suspect that either $i$ and $j$ are members of the same almost invariant aggregate $\mathcal{E}_k$ or $i \in \mathcal{E}_0$.

We present Algorithm 5, which will be fleshed out into three implementable versions. The input for this base code is a stochastic matrix $A$ on the state space $\mathcal{S}$.

Now, suppose that $A$ is a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$

**Algorithm 5** Error-reducing base code

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that contains no arcs.
$\mathcal{C} := \emptyset$
**while** the order of $B$ is 2 or greater **do**
  Select $i \in \mathcal{S} \setminus \mathcal{C}$ such that the stochastic complement $B \setminus i$ is error-reducing with respect to some $\epsilon$-uncoupling of $A$.
  **if** no such $i \in \mathcal{S} \setminus \mathcal{C}$ exists **then**
    Exit the **while** loop.
  **else**
    $\mathcal{C} := \mathcal{C} \cup \{i\}$
    Select $j \in \mathcal{S} \setminus \mathcal{C}$ such that $b_{ij} = \max_{j' \in \mathcal{S} \setminus \mathcal{C}} \{b_{i'j'}\}$.
    Add the directed arc $i \to j$ to $G$.
    $B := B \setminus i$
  **end if**
**end while**
**return** $G$

and that the digraph $G$ has been constructed via an application of an error-reducing algorithm. Then, whenever the directed arc $i \to j$ is present in $G$, there is an error-reducing complement $\hat{A}$ of $A$ where the entry $\hat{a}_{ij}$ is the largest off-diagonal entry in the $i$th row of $\hat{A}$. As discussed above, we suspect that $\hat{a}_{ij}$ is not an error term (a transition between almost invariant aggregates), but represents a regularly occurring transition. Thus, if $i \prec_G j$, we suspect that either $i$ is near transient or that $i$ and $j$ are contained in the same almost invariant aggregate. As in Proposition 5.4, the weakly connected components of the output digraph $G$ are strong candidates for almost invariant aggregates of the Markov chain associated with $A$.

### 6.1.2 Diagonal bounds

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$, let $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$ be an $\epsilon$-uncoupling and let $A \setminus \mathcal{C}$ be an error-reducing stochastic complement. Suppose that we have associated each state $i \in \mathcal{C}$ with a unique state $j \in \mathcal{S} \setminus \mathcal{C}$, with the association denoted by $i \sim j$. Suppose further that if $i \sim j$ where $i \in \mathcal{C}$ and $j \in \mathcal{S} \setminus \mathcal{C}$, then either $i \in \mathcal{E}_0$ or $i, j \in \mathcal{E}_k$ for some $k \neq 0$. For each $j \in \mathcal{S} \setminus \mathcal{C}$, let

$$m_j = |\{i \in \mathcal{C} : i \sim j\}| + 1.$$

Suppose that there is $k \neq 0$ such that $\mathcal{E}_k \setminus \mathcal{C} = \{i\}$. We note that $m_i \geq |\mathcal{E}_k|$. Now, because the stochastic complement is error reducing, and $\hat{A}(\mathcal{E}_k \setminus \mathcal{C}) = \hat{a}_{ii}$, we have

$$1 - \hat{a}_{ii} \leq \eta(A(\mathcal{E}_k)) \leq |\mathcal{E}_k|\epsilon \leq m_i\epsilon.$$

Thus, $\hat{a}_{ii} \geq 1 - m_i\epsilon$.

So, we may use the following as test when applying Algorithm 5 to determine whether or not a given $i \in \mathcal{S} \setminus \mathcal{C}$ is the final unremoved member of an almost invariant aggregate.

1. If $\hat{a}_{ii} \geq 1 - m_i\epsilon$, state $i$ may be the final unremoved member of an almost invariant aggregate, in which case the stochastic complement $\hat{A} \setminus i$ is not error reducing.

2. If $\hat{a}_{ii} < 1 - m_i\epsilon$, state $i$ is a candidate for removal.

We note that in case 2, above, we need to select $i$ carefully to ensure that $\hat{A} \setminus i$ is also error-reducing – the fact that $\hat{a}_{ii}$ is not close to 1 is not, in itself, sufficient to ensure that state $i$ is safe to remove.

However, we supsect that the criterion $\hat{a}_{ii} < 1 - m_i\epsilon$ (to determine if $i$ is safe to remove) is, in general, too conservative. For example, if $\mathcal{E}$ is an almost invariant aggregate such that $|\mathcal{E}|\epsilon > 1$, it is impossible for an implementation that uses this criterion to correctly associate all the members of $\mathcal{E}$.

We instead use the bounds calculated in Appendices A and B (the bound in Appendix B has already been utilised for the Modified Maximum Entry Algorithm).

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$ and let $\hat{A} = A \setminus \mathcal{C}$ be a stochastic complement where we have associated each member of $\mathcal{C}$ with a unique member of $\mathcal{S} \setminus \mathcal{C}$. Let $i \in \mathcal{S} \setminus \mathcal{C}$ and let $m_i - 1$ be the number of states contained in $\mathcal{C}$ which have been associated with $i$. Then,

1. if $A$ is reversible and

$$\hat{a}_{ii} < \frac{(1-\epsilon)^2}{1 + (m_i - 2)\epsilon},$$

the state $i$ is a candidate for removal; and

2. if $A$ is nonreversible and

$$\hat{a}_{ii} < (1-\epsilon)^{m_i},$$

the state $i$ is a candidate for removal.

**Proposition 6.1.** *Let $m$ be a positive integer and let $0 \leq \epsilon < 1$. Then,*

$$1 - m\epsilon \leq (1-\epsilon)^m \leq \frac{(1-\epsilon)^2}{1+(m-2)\epsilon} \leq 1.$$

*Moreover,*

1. *$1 - m\epsilon = (1-\epsilon)^m$ if and only if $m = 1$ or $\epsilon = 0$,*
2. *$(1-\epsilon)^m = \frac{(1-\epsilon)^2}{1+(m-2)\epsilon}$ if and only if $m = 1$, $m = 2$, or $\epsilon = 0$, and*
3. *$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = 1$ if and only if $\epsilon = 0$.*

**Proof** Clearly, for $\epsilon = 0$ we have equality of all four terms involved; so, we assume that $0 < \epsilon < 1$.

We first show that $1 - m\epsilon \leq (1-\epsilon)^m$ with equality if and only if $m = 1$. We proceed by induction on $m$. For $m = 1$, $1 - m\epsilon = (1-\epsilon)^m = 1 - \epsilon$. Suppose that $m \geq 1$ and that $1 - m\epsilon \leq (1-\epsilon)^m$. Then,

$$\begin{aligned}
(1-\epsilon)^{m+1} &= (1-\epsilon)(1-\epsilon)^m \\
&\geq (1-\epsilon)(1-m\epsilon) \\
&= 1 - (m+1)\epsilon + m\epsilon^2 \\
&> 1 - (m+1)\epsilon.
\end{aligned}$$

We now show that

$$(1-\epsilon)^m \leq \frac{(1-\epsilon)^2}{1+(m-2)\epsilon},$$

with equality if and only if $m = 1$ or $m = 2$. For $m = 1$

$$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = \frac{(1-\epsilon)^2}{1-\epsilon} = 1-\epsilon,$$

and for $m = 2$

$$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = (1-\epsilon)^2.$$

So, assume that $m \geq 3$. We note that

$$1 - \frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon}.$$

So, we will prove that

$$1-(1-\epsilon)^m > 1 - \frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = \frac{\epsilon(m-\epsilon)}{1+(m-2)\epsilon},$$

by proving that

$$\frac{\left(1-(1-\epsilon)^m\right)\left(1+(m-2)\epsilon\right)}{\epsilon} > m-\epsilon.$$

We make use of the well-known formula $1 - z^m = (1-z)(1+z+\cdots+z^{m-1})$. This, together with the facts that $m \geq 3$ and $0 < 1-\epsilon < 1$, implies that

$$\begin{aligned}
\tfrac{1-(1-\epsilon)^m}{\epsilon} &= \tfrac{(1-(1-\epsilon))\left(1+(1-\epsilon)+\cdots+(1-\epsilon)^{m-1}\right)}{\epsilon} \\
&= 1+(1-\epsilon)+\cdots+(1-\epsilon)^{m-1} \\
&> 1+(1-\epsilon)+(m-2)(1-\epsilon)^m.
\end{aligned}$$

So,

$$\begin{aligned}
\tfrac{(1-(1-\epsilon)^m)(1+(m-2)\epsilon)}{\epsilon} &= \tfrac{(1-(1-\epsilon)^m)}{\epsilon} + (m-2)\left(1-(1-\epsilon)^m\right) \\
&> 1 + (1-\epsilon) + (m-2)(1-\epsilon)^m + (m-2)\left(1-(1-\epsilon)^m\right) \\
&= m - \epsilon.
\end{aligned}$$

Finally, we note that $(1-\epsilon)^2 < 1-\epsilon$ and $1+(m-2)\epsilon \geq 1-\epsilon$ imply that

$$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon} < 1.$$

∎

In the analysis of power series and other continuous functions, big $\Theta$ notation is used to describe the behaviours of functions as they approach a specific limit. Let $f$ and $g$ be real functions and let $\alpha$ be a real number. We say that

$$f(x) = \Theta(g(x)) \text{ as } x \to \alpha$$

if there are positive constants $c$, $d$ and $\delta$ such that if $|x-\alpha| < \delta$, then

$$cg(x) \leq f(x) \leq dg(x).$$

The notation

$$f(x) = g(x) + \Theta(h(x))$$

is used to represent the fact that

$$f(x) - g(x) = \Theta(h(x)) \text{ as } x \to \alpha.$$

When $f(x) = g(x) + \Theta(h(x))$ as $x \to \alpha$ and

$$\lim_{x \to \alpha} h(x) = 0,$$

the function $g(x)$ is seen to be a good approximation of $f(x)$, near $x = \alpha$. We use the same big $\Theta$ notation for functions on vector spaces.

In addition to the results in Proposition 6.1, we note that for $m \geq 1$ and $0 < \epsilon < 1$,

$$(1-\epsilon)^m = 1 - m\epsilon + \binom{m}{2}\epsilon^2 - \ldots + (-\epsilon)^m = 1 - m\epsilon + \Theta(m^2\epsilon^2)$$

and

$$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon} = 1 - m\epsilon + \frac{(m-1)^2\epsilon^2}{1+(m-2)\epsilon} = 1 - m\epsilon + \Theta(m^2\epsilon^2)$$

as $m\epsilon \to 0$. Thus, if the value $n^2\epsilon^2$ is insignificant (where $n$ is the order of the stochastic matrix in question), the three criteria presented here are substantially the same.

## 6.2 The Lower Weighted Algorithm

We present an algorithm, intended for use with reversible stochastic matrices, which attempts to construct almost invariant aggregates of a reversible Markov chain via subsequent error-reducing complements.

### 6.2.1 Reordering reversible stochastic matrices

Let $A$ be a reversible substochastic matrix on the state space $\mathcal{S}$. Throughout the remainder of this section, we will typically assume that $\mathcal{S} = \{1, \ldots, n\}$ for some $n \geq 1$. This allows us to assume that there is a natural, transitive ordering of the set $\mathcal{S}$, denoted by the symbol $<$.

Let $(k_1, \ldots, k_m)$ be a sequence of distinct members of $\mathcal{S}$. We define $A(k_1, \ldots, k_m)$ to be the $m \times m$ substochastic matrix on the state space $\{1, \ldots, m\}$ whose $ij$th entry is equal to $a_{k_i k_j}$. That is, $A(k_1, \ldots, k_m)$ is the principal submatrix of $A$ corresponding to the collection $\{k_1, \ldots, k_m\}$ and the states have been reordered via the sequence indices of $(k_1, \ldots, k_m)$. If $A$ is $m \times m$, we refer to $A(k_1, \ldots, k_m)$ as a reordering of $A$.

Let $A$ be a reversible substochastic matrix on the ordered state space $\mathcal{S}$ (typically $\mathcal{S}$ is some finite subset of the positive integers). We refer to $A$ as *lower-weighted* if for all $i, j \in \mathcal{S}$ with $i < j$, we have $a_{ij} \leq a_{ji}$.

As usual, we use the abbreviation $\pi_i$ to represent the $i$th diagonal entry of a diagonal matrix $\Pi$.

**Proposition 6.2.** *Let $A$ be a reversible substochastic matrix on the ordered state space $\mathcal{S}$. Then, there is a reordering of $A$ that is lower-weighted.*

**Proof** Without loss of generality, we assume that $\mathcal{S} = \{1, \ldots, n\}$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Let $f : \mathcal{S} \mapsto \mathcal{S}$ be a permutation such that if $i < j$, then $\pi_{f(i)} \geq \pi_{f(j)}$. Via Proposition 2.12,

$$\pi_i a_{ij} = \pi_j a_{ji},$$

for all $i, j \in \mathcal{S}$. If $\pi_i \geq \pi_j$, we have $a_{ij} \leq a_{ji}$. Thus, for all $i, j \in \mathcal{S}$, if $i < j$, then $\pi_{f(i)} \geq \pi_{f(j)}$ and $a_{f(i)f(j)} \leq a_{f(j)f(i)}$. So, the matrix

$$A(f(1), \ldots, f(n))$$

is a lower-weighted reordering of $A$. ∎

The Reorder Algorithm produces a lower-weighted reordering of an input reversible substochastic matrix. The input of Algorithm 6 is a reversible substochastic matrix $A$ on the state space $\{1, \ldots, n\}$; the output is an ordering of the states into $(f(1), \ldots, f(n))$ such that $A(f(1), \ldots, f(n))$ is lower-weighted. We assume that $n \geq 2$, as we consider any $1 \times 1$ stochastic matrix to be lower-weighted, trivially.

When the substochastic matrix $A$ has state space $\mathcal{S} = \{i_1, \ldots, i_n\} \neq \{1, \ldots, n\}$, the reorder algorithm can still be applied. One must simply replace the opening command $f := (1, \ldots, n)$ with $f := (i_1, \ldots, i_n)$.

We have based the Reorder Algorithm on a graph searching algorithm known as Depth-First Search [5, Section 22.3].

**Lemma 6.3.** *Let $A$ be a reversible substochastic matrix on the states $\{1, \ldots, n\}$ where $n \geq 2$. Then, Algorithm 6, applied to $A$, will terminate after a finite number of*

**Algorithm 6** Reorder

$f := (1, \ldots, n)$
$r := 1$
$s := 2$
**while** $s \leq n$ **do**
  **if** $a_{f(r)f(t)} \leq a_{f(t)f(r)}$ for $t = s, \ldots, n$ **then**
    $r := r + 1$
    **if** $r = s$ **then**
      $s := s + 1$
    **end if**
  **else**
    Let $t$ be such that $s \leq t \leq n$ and $a_{f(r)f(t)} > a_{f(t)f(r)}$.
    $(f(r), f(r+1), \ldots, f(t)) := (f(t), f(r), f(r+1), \ldots, f(t-1)$
    $s := s + 1$
  **end if**
**end while**
**return** $f$

*iterations of its internal* ***while*** *loop. Furthermore, after any number of iterations of the* ***while*** *loop, the stored data* $r$ *and* $s$ *satisfies* $1 \leq r < s$.

**Proof** The algorithm begins with $r = 1$ and $s = 2$. Every iteration of the algorithm increases one or both of $r$ and $s$ by 1. Thus, at any iteration $r \geq 1$, $s \geq 2$ and $r$ and $s$ are positive integers. The **while** continues only if $s \leq n$, so the algorithm terminates only if it achieves $s = n + 1$.

We see that at any point $r < s$ because the initial data $r = 1$ and $s = 2$ has $r < s$, and whenever the algorithm encounters the command $r := r + 1$, this is immediately followed by the command **if** $r = s$ **then** $s := s + 1$.

As we noted above, each iteration of the while loop increases one or both of $r$ and $s$ by 1; thus, after $k$ iterations of the **while** loop, $r + s \geq 3 + k$. So, after $k = 2n - 3$

iterations, since $r < s$, we have

$$2s > r + s \geq 3 + k = 2n,$$

further implying that $s \geq n+1$. Thus, the algorithm terminates after $2n-3$ or fewer iterations. ∎

**Proposition 6.4.** *Let $A$ be a reversible substochastic matrix of order $n \geq 2$ and suppose that we have applied Algorithm 6 to $A$. Then, the output $f$ corresponds to a lower-weighted reordering of $A$.*

**Proof** Each iteration of the algorithm either leaves the sequence $f$ fixed or it permutes a subsequence of $f$ (before altering the values of the stored variables $r$ and $s$):

$$(f(r), f(r+1), \ldots, f(t)) := (f(t), f(r), f(r+1), \ldots, f(t-1)).$$

When this operation occurs we have $t \geq s$; by Lemma 6.3, we always have $r < s$, so this is indeed a permutation of $f$. Thus, after any number of iterations of the **while** loop of Algorithm 6, the stored sequence $f$ is a permutation of the initial sequence $(1, \ldots, n)$.

Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Via Lemma 6.3, after any number of iterations of the **while** loop, we have $1 \leq r < s$. Moreover, either $s \leq n$ or $s = n+1$ and the algorithm terminates without executing another iteration.

We will show, by induction on the number of completed iterations of the **while** loop, that the stored data $r$, $s$ and $f$ (after any number of iterations) satisfies each of the following statements:

1. If $1 \leq i < j < s$, then $a_{f(i)f(j)} \leq a_{f(j)f(i)}$.

2. If $1 \leq i < r$ and $i < j \leq n$, then $a_{f(i)f(j)} \leq a_{f(j)f(i)}$.

3. The values $\pi_{f(r)} > \pi_{f(r+1)} > \ldots > \pi_{f(s-1)}$ are strictly decreasing.

The algorithm terminates with $s = n + 1$; thus, at termination, statement 1 implies that $A(f(1), \ldots, f(n))$ is a lower-weighted reordering of $A$. (The other two statements are necessary for the inductive reasoning.)

At initialisation we have $r = 1$ and $s = 2$; there are no integer values $i$ and/or $j$ which satisfy $1 \leq i < j < s = 2$ or $1 \leq i < r = 1$, so the first two statements trivially hold. Moreover, $r = s - 1$ at initialisation, so the sequence in the fourth statement contains one element and so is (trivially) strictly decreasing.

Now, suppose that the data $r$, $s$, and $f$ satisfies the three statements above; suppose further that $s \leq n$ and let $r'$, $s'$ and $f'$ be the new stored data after one further iteration of the **while** loop. We will show that the three statements above hold for $r'$, $s'$ and $f'$. We will refer to the values $r$, $s$ and $f$ as the previous data and the values $r'$, $s'$ and $f'$ as the current data.

*Case one:* $a_{f(r)f(t)} \leq a_{f(t)f(r)}$ for $t := s, \ldots, n$.

In this case, we have $r' = r + 1$, $f' = f$, and either $r < s - 1$ and $s' = s$, or $r = s - 1$ and $s' = s + 1$.

We first show that statement 2 holds true for the new data. Suppose that $1 \leq i < r'$ and that $i < j \leq n$. Since $r' = r + 1$ we either have $i < r$ or $i = r$. If $i < r$, then $a_{f(i)f(j)} \leq a_{f(j)f(i)}$, as statement 2 holds true for the previous data. So, we need merely show that if $r < j \leq n$, then $a_{f(r)f(j)} \leq a_{f(j)f(r)}$.

First, suppose that $j \leq s - 1$. Then, statement 1, applied to the previous data, implies that $a_{f(r)f(j)} \leq a_{f(j)f(r)}$. Secondly, if $j \geq s$, the fact that $a_{f(r)f(j)} \leq a_{f(j)f(r)}$ is the base assumption of this case.

Now, we show that statements 1 and 3 hold true for the current data.

Suppose that $r < s - 1$; then, $s' = s$. Thus, statement 1 holds true for the current data, as it holds true for the previous data and only concerns $s' = s$ and $f' = f$. The third statement holds true for the current data because the sequence $(f(r'), \ldots, f(s'-1)) = (f(r+1), \ldots, f(s-1))$ is a subsequence of $(f(r), \ldots, f(s-1))$.

Suppose that $r = s-1$; then, $s' = s+1$. In this case, statement 1 is a consequence of statement 2 (which we have shown to hold for the current data): if $1 \leq i < j < s'$, then $i < s' - 1 = r'$, so we have $a_{f(i)f(j)} \leq a_{f(j)f(i)}$. As well, the sequence involved in statement 3, $\pi_{f(r')}, \ldots, \pi_{f(s'-1)}$, contains only one element, since $r' = s = s' - 1$, and so is strictly decreasing.

*Case two:* The algorithm has selected an index $t$ is such that $s \leq t \leq n$ and

$a_{f(r)f(t)} > a_{f(t)f(r)}$.

In this case $r' = r$, $s' = s + 1$,

$$\begin{aligned} f'(r) &= f(t), \\ f'(r+1) &= f(r), \\ f'(r+2) &= f(r+1), \\ &\vdots \\ f'(t) &= f(t-1), \end{aligned}$$

and for $i < r$ or $i > t$, $f'(i) = f(i)$.

We first show that statements 2 and 3 hold for the current data.

Statement 2 is the claim that if $1 \leq i < r'$ and $i < j \leq n$, then $a_{f'(i)f'(j)} \leq a_{f'(j)f'(i)}$. Let $1 \leq i < r'$ and $i < j \leq n$. Note that since $r' = r$, $f'(i) = f(i)$. First, suppose that $j < r$; then $f'(j) = f(j)$. Thus, since statement 2 holds for the previous data, we have

$$a_{f'(i)f'(j)} = a_{f(i)f(j)} \leq a_{f(j)f(i)} = a_{f'(j)f'(i)}.$$

Second, suppose that $j \geq r$; since the permutation that transforms $f$ into $f'$ fixes the first $r - 1$ elements, there is $j' \geq r$ such that $f'(j) = f(j')$. Again, the fact that statement 2 holds for the previous data implies that

$$a_{f'(i)f'(j)} = a_{f(i)f(j')} \leq a_{f(j')f(i)} = a_{f'(j)f'(i)}.$$

The sequence in question in statement 3 is

$$(f'(r'), f'(r'+1), \ldots, f'(s'-1)) = (f(t), f(r), f(r+1), \ldots, f(s-1)).$$

Since statement 3 applies to the previous data, we have $\pi_{f(r)} > \ldots > \pi_{f(s-1)}$. Thus, we only need to show that $\pi_{f(t)} > \pi_{f(r)}$. Since $\Pi A$ is symmetric, this is a direct consequence of the fact that $a_{f(r)f(t)} > a_{f(t)f(r)}$.

Now, we show that statement 1 holds true for the current data. That is, we show that if $1 \leq i < j < s'$, then $a_{f'(i)f'(j)} \leq a_{f'(j)f'(i)}$. If $i < r'$, this is true via the fact that statement 2 holds for the current data. If $i \geq r'$, then the fact that statement 3 holds for the current data implies that $\pi_{f'(i)} > \pi_{f'(j)}$; this in turn implies that $a_{f'(i)f'(j)} \leq a_{f'(j)f'(i)}$ (again, since $A$ is reversible). ∎

### 6.2.2 Error reduction in lower-weighted matrices

We explore the effect that removing states from reversible stochastic matrices can have on the error values of almost invariant aggregates.

**Proposition 6.5.** *Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$. Let $i \in \mathcal{S} \setminus \mathcal{E}$ be such that the stochastic complement $\hat{A} = A \setminus i$ exists. Let $B = A(\mathcal{E})$ and $\hat{B} = \hat{A}(\mathcal{E})$. Then, $\eta(\hat{B}) \leq \eta(B)$.*

**Proof** For $i', j' \in \mathcal{S} \setminus i$,

$$\hat{a}_{i'j'} = a_{i'j'} + \frac{a_{i'i}a_{ij'}}{1 - a_{ii}} \geq a_{i'j'}.$$

Thus, since $i \notin \mathcal{E}$, $\hat{B} \geq B$, further implying that

$$\eta(\hat{B}) = \mathbb{1}^T(I - \hat{B})\mathbb{1} \leq \mathbb{1}^T(I - B)\mathbb{1} = \eta(B).$$

∎

**Proposition 6.6.** *Let $A$ be a reversible stochastic matrix on the state space $\mathcal{S} = \{1, \ldots, n\}$, let $\mathcal{E} \subseteq \mathcal{S}$ contain 2 or more states and let $B = A(\mathcal{E})$. Suppose that $A(f(1), \ldots, f(n))$ is a lower-weighted reordering of $A$ and let*

$$k = \max_{1 \leq k' \leq n} \{k' : f(k') \in \mathcal{E}\}.$$

*Suppose further that the stochastic complement $\hat{A} = A \setminus f(k)$ exists and let $\hat{B} = B \setminus f(k) = \hat{A}(\mathcal{E} \setminus f(k))$. Then, $\eta(\hat{B}) \leq \eta(B)$.*

**Proof** Without loss of generality, we will show that the result holds for $A$ lower-weighted (the function $f$ above is the identity).

Express $\mathcal{E} = \{k_1, \ldots, k_m\}$ where $m \geq 2$ and $k_1 < \ldots < k_m$; then

$$k_m = \max_{1 \leq k' \leq n} \{k' : k' \in \mathcal{E}\}.$$

Let $i = k_m$ and express

$$A(\mathbb{E}) = B = \begin{bmatrix} \tilde{B} & v \\ w^T & b_{ii} \end{bmatrix},$$

where the final position corresponds to state $i = k_m$. Because $A$ is lower-weighted, $B$ is lower-weighted, as well. This implies that $v \leq w$. We calculate

$$\eta(B) = \mathbb{1}^T(I - B)\mathbb{1} = m - \mathbb{1}^T\tilde{B}\mathbb{1} - \mathbb{1}^T v - w^T\mathbb{1} - b_{ii},$$

$$\hat{B} = B \setminus i = \tilde{B} + \frac{1}{1 - b_{ii}} v w^T$$

and

$$\eta(\hat{B}) = \mathbb{1}^T(I - \hat{B})\mathbb{1} = m - 1 - \mathbb{1}^T\tilde{B}\mathbb{1} - \frac{1}{1 - b_{ii}}\mathbb{1}^T v w^T \mathbb{1}.$$

We aim to show that $\eta(\hat{B}) \leq \eta(B)$, which occurs only if

$$\mathbb{1}^T v - \frac{\mathbb{1}^T v w^T \mathbb{1}}{1 - b_{ii}} \leq 1 - b_{ii} - w^T\mathbb{1}.$$

Now, since $w^T\mathbb{1} \leq 1 - b_{ii}$ ($B$ is substochastic) and $v \leq w$,

$$\frac{\mathbb{1}^T v}{1 - b_{ii}} \leq \frac{\mathbb{1}^T w}{1 - b_{ii}} \leq 1.$$

Therefore,

$$\mathbb{1}^T v - \frac{\mathbb{1}^T v w^T \mathbb{1}}{1 - b_{ii}} = \frac{\mathbb{1}^T v}{1 - b_{ii}} \left(1 - b_{ii} - w^T\mathbb{1}\right) \leq 1 - b_{ii} - w^T\mathbb{1}.$$

∎

**Corollary 6.7.** *Let $A$ be a nearly uncoupled reversible stochastic matrix on the state space $\mathcal{S}$ and let $\hat{A} = A \setminus \mathcal{C}$ be an error-reducing complement, with respect to some $\epsilon$-uncoupling $\Psi$. Let $\hat{A}(f(1), \ldots, f(n))$ be a lower-weighted reordering of $\hat{A}$ and let $k$*

*be the largest index less than or equal $n$ such that $\mathcal{C} \cup \{f(k)\}$ does not contain an almost invariant aggregate. Then, $\hat{A} \setminus f(k)$ is an error-reducing complement, with respect to $\Psi$.*

**Proof** Let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate in the uncoupling $\Psi$ and let $\mathcal{E}' = \mathcal{E} \setminus \mathcal{C}$; since $A \setminus \mathcal{C}$ is error-reducing $\mathcal{E}'$ is nonempty. Let $\hat{A}$, $f$ and $k$ be as described above and let

$$\tilde{A} = \hat{A} \setminus f(k) = A \setminus (\mathcal{C} \cup \{f(k)\}).$$

First, suppose that $f(k) \notin \mathcal{E}$. Let $B = A(\mathcal{E})$, $\hat{B} = \hat{A}(\mathcal{E}')$ and $\tilde{B} = \tilde{A}(\mathcal{E}')$ (since $f(k) \notin \mathcal{E}$, $\mathcal{E} \setminus (\mathcal{C} \cup \{f(k)\} = \mathcal{E} \setminus \mathcal{C} = \mathcal{E}'$). By Proposition 6.4, $\eta(\tilde{B}) \leq \eta(\hat{B})$ and, since $A \setminus \mathcal{C}$ is error-reducing, $\eta(\hat{B}) \leq \eta(B)$.

Now, suppose that $f(k) \in \mathcal{E}$; since $\mathcal{E} \not\subseteq \mathcal{C} \cup \{f(k)\}$ (by assumption), $\mathcal{E}' = \mathcal{E} \setminus \mathcal{C}$ contains two or more states. Let $\mathcal{E}'' = \mathcal{E}' \setminus f(k)$. For any $k' \neq k$ such that $f(k') \in \mathcal{E}$, we then have $\mathcal{E} \not\subseteq \mathcal{C} \cup \{f(k')\}$, implying, via the definition of $k$, that $k' < k$. Thus, $k$ is the largest integer less than or equal to $n$ such that $f(k) \in \mathcal{E}$. Let $B = A(\mathcal{E})$, $\hat{B} = \hat{A}(\mathcal{E}')$ and $\tilde{B} = \tilde{A}(\mathcal{E}'')$. By Proposition 6.5, $\eta(\tilde{B}) \leq \eta(\hat{B})$ and, as above, $\eta(\hat{B}) \leq \eta(B)$.

Thus, for any almost invariant aggregate $\mathcal{E}$ in $\Psi$, the error-inflation at $\mathcal{E}$ induced by removing $f(k)$ is less than or equal to 1. ∎

### 6.2.3 The Lower-Weighted Algorithm

**Algorithm 7** Choose

$p := \max_{j \neq i}\{a_{ij}\}$
$K := \{j \neq i : a_{ij} = p\}$
$q := \min_{j \in K}\{a_{ji}\}$
**if** $q > p$ or $p = 0$ **then**
  **return** 0
**else**
  Choose a state $j \in K$ that has $a_{ji} = q$.
  **return** $j$
**end if**

We present the Lower-Weighted Algorithm, which attempts to construct almost invariant aggregates of a given reversible stochastic matrix. We first present a sub-algorithm, which will be of use in the main pseudocode, below. The inputs of the Choose Algorithm are a reversible stochastic matrix $A$ on the state space $\mathcal{S}$ and a single state $i \in \mathcal{S}$. The Choose Algorithm implicitly assumes that $0 \notin \mathcal{S}$. If $0 \in \mathcal{S}$, we need to utilise some other symbol, not contained in $\mathcal{S}$, in its place. The output of the Choose Algorithm is

1. a state $j \in \mathcal{S}$, distinct from $i$, such that

$$a_{ij} = \max_{j' \neq i}\{a_{ij'}\}$$

   and $\pi_j \geq \pi_i$ for any stationary distribution $\pi$ of $A$, or

2. 0, if no such $j$ exists.

The inputs of the Lower-Weighted Algorithm are a stochastic matrix $A$ on the state space $\mathcal{S} = \{1, \ldots, n\}$ and a small nonnegative value $\epsilon < 1$.

**Algorithm 8** The lower-Weighted Algorithm

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that contains no arcs.
$m := \mathbb{1}_{\mathcal{S}}$
$f := \text{Reorder}(B)$
$n := |\mathcal{S}|$
**while** $n \geq 2$ **do**
  **if** for $k = 1, \ldots, n$, $b_{f(k)f(k)} \geq \frac{(1-\epsilon)^2}{1+(m_{f(k)}-2)\epsilon}$ or $\text{Choose}(B, f(k)) = 0$ **then**
    Exit the **while** loop.
  **else**
    $k := \max_{1 \leq k' \leq n} \{k' : b_{f(k')f(k')} < \frac{(1-\epsilon)^2}{1+(m_{f(k')}-2)\epsilon} \text{ and } \text{Choose}(B, f(k')) \neq 0\}$
    $L := \{(l_1, l_2) : l_1 \neq l_2, \ b_{f(l_1)f(l_2)} = 0 \text{ and } b_{f(l_1)f(k)} a_{f(k)f(l_2)} \neq 0\}$
    $j := \text{Choose}(B, f(k))$
    Add the directed arc $f(k) \to j$ to $G$.
    $m_j := m_j + m_{f(k)}$
    $B := B \setminus f(k)$
    $f := (f(1), \ldots, f(k-1), f(k+1), \ldots, f(n))$
    $n := n - 1$
    **if** $L$ is nonempty **then**
      $l_{\min} := \min_{(l_1, l_2) \in L} \{l_1\}$
      **if** $l_{\min} > k$ **then**
        $l_{\min} := l_{\min} - 1$
      **end if**
      $l_{\max} := \max_{(l_1, l_2) \in L} \{l_1\}$
      **if** $l_{\max} > k$ **then**
        $l_{\max} := l_{\max} - 1$
      **end if**
      $g := \text{Reorder}(B(f(l_{\min}), \ldots, f(l_{\max})))$
      $(f(l_{\min}), \ldots, f(l_{\max})) := (f(l_{\min} - 1 + g(1)), \ldots, f(l_{\min} - 1 + g(l_{\max} - l_{\min} + 1))$
    **end if**
  **end if**
**end while**
**return** $G$

**Proposition 6.8.** *Let $A$ be a reversible stochastic matrix and suppose that Algorithm 8 has been applied to $A$. Let $B$, $f$, $l$, $G$ and $m$ be the stored data after any number of iterations of the algorithm's* ***while*** *loop. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric, let $F = \{f(1), \ldots, f(l)\}$ and let $\mathcal{C} = \mathcal{S} \setminus F$. Then,*

1. *$B = A \setminus \mathcal{C}$;*
2. *$B(f(1), \ldots, f(l))$ is a lower-weighted reordering of $B$;*
3. *$G$ is acyclic, every member of $\mathcal{C}$ has out-degree $1$ in $G$ and every member of $F$ has out-degree $0$;*
4. *if the directed arc $i \to j$ is present in $G$, then $\pi_i \leq \pi_j$; and*
5. *for each $i \in F$, the weakly connected component of $G$ containing $i$ contains exactly $m_i$ states.*

**Proof** Statements 1, 3 and 5 are shown in the same manner as in Proposition 5.1 and Lemma 5.3.

Statement 4 is a consequence of the workings of the Choose Algorithm. Suppose that the directed arc $i \to j$ is present in $G$; then, there is a stochastic complement $B = A \setminus \mathcal{C}$ such that $\text{Choose}(B, i) = j$, which implies that $\pi(j) \geq \pi(i)$.

We now show statement 2. In Proposition 6.4, we have shown that the reorder algorithm produces a lower-weighted reordering of a reversible stochastic matrix; thus,

the matrix $B$ and the permutation $f := \text{Reorder}(B)$, at initialisation, satisfy statement 2. We show that if $f$ corresponds to a lower-weighted reordering of $B$, one further iteration of the algorithm does not alter this fact.

Let $B$, $f$ and $l$ be the stored data after some number of iterations and suppose that $B(f(1), \ldots, f(n))$ is a lower-weighted reordering of $B$. Suppose further that the algorithm executes at least one more iteration before terminating and let $B'$, $f'$ and $n' = n - 1$ be the stored data after one more iteration. Let $f(k)$ be the state selected for removal and let

$$L := \{(l_1, l_2) : l_1 \neq l_2, \quad b_{f(l_1)f(l_2)} = 0 \quad \text{and} \quad b_{f(l_1)f(k)} a_{f(k)f(l_2)} \neq 0\}.$$

*Case one:* $L$ is empty.

Then, we have

$$f' = (f(1), \ldots, f(k-1), f(k+1), \ldots, f(n)).$$

So, we need to show that if $1 \leq i < j \leq n$ and $i, j \neq k$, then $b'_{f(i)f(j)} \leq b'_{f(j)f(i)}$. We note that

$$b'_{f(i)f(j)} = \frac{b_{f(i)f(k)} b_{f(k)f(j)}}{1 - b_{f(k)f(k)}} \quad \text{and} \quad b'_{f(j)f(i)} = \frac{b_{f(j)f(k)} b_{f(k)f(i)}}{1 - b_{f(k)f(k)}}.$$

Let $1 \leq i < j \leq n$ and $i, j \neq k$. Since $L$ is empty, we have either $b_{f(i)f(j)} \neq 0$ or $b_{f(i)f(k)} b_{f(k)f(j)} = 0$.

Suppose that $b_{f(i)f(j)} \neq 0$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric. Then, via Propositions 2.12 and 4.6,

$$\pi_{f(i)}b_{f(i)f(j)} = \pi_{f(j)}b_{f(j)f(i)} \quad \text{and} \quad \pi_{f(i)}b'_{f(i)f(j)} = \pi_{f(j)}b'_{f(j)f(i)}.$$

Since $b_{f(i)f(j)} \leq b_{f(j)f(i)}$ and $b_{f(i)f(j)} \neq 0$, we must have $\pi_{f(i)} \geq \pi_{f(j)}$, which in turn implies that $b'_{f(i)f(j)} \leq b'_{f(j)f(i)}$.

Suppose that $b_{f(i)f(k)}b_{f(k)f(j)} = 0$. Then, either $b_{f(i)f(k)} = 0$ or $b_{f(k)f(j)} = 0$, implying (as $B$ is reversible) that either $b_{f(k)f(i)} = 0$ or $b_{f(j)f(k)} = 0$. Thus, $b'_{f(i)f(j)} = b_{f(i)f(j)}$ and $b'_{f(j)f(i)} = b_{f(j)f(i)}$. So, since $b_{f(i)f(j)} \leq b_{f(j)f(i)}$, we have $b'_{f(i)f(j)} \leq b'_{f(j)f(i)}$.

*Case two:* $L$ is nonempty.

Let

$$L_{\min} = \min_{(l_1,l_2)\in L}\{l_1\} \quad \text{and} \quad L_{\max} = \max_{(l_1,l_2)\in L}\{l_1\};$$

let

$$l_{\min} = \begin{cases} L_{\min} & \text{if } L_{\min} < k \\ L_{\min} - 1 & \text{otherwise,} \end{cases} \quad \text{and} \quad l_{\max} = \begin{cases} L_{\max} & \text{if } L_{\max} < k \\ L_{\max} - 1 & \text{otherwise.} \end{cases}$$

We note that $(k, l') \notin L$, for any index $l'$. If we suppose that $(k, l') \in L$, then

$$b_{f(k)f(l')} = 0 \quad \text{and} \quad b_{f(k)f(k)}b_{f(k)f(l')} \neq 0,$$

which is a contradiction. We further note that there are no elements of the form $(l', l')$ contained in $L$; so, $l_{\min} < l_{\max}$. Let $g = \text{Reorder}(B'(l_{\min}, \ldots, l_{\max}))$; we note that $g$ is a permutation of the indices $1, \ldots, l_{\max} - l_{\min} + 1$.

The permutation $f'$ is formed by first removing the $k$th element of $f$, forming

$$\hat{f} = (f(1), \ldots, f(k-1), f(k+1), \ldots, f(l)),$$

and then permuting the subsequence consisting of the $l_{\min}$th through $l_{\max}$th elements,

$$f' = (\hat{f}(1), \ldots, \hat{f}(l_{\min}-1), f'(l_{\min}), \ldots, f'(l_{\max}), \hat{f}(l_{\max}+1), \ldots, \hat{f}(n-1)),$$

where

$$\begin{aligned} f'(l_{\min}) &= \hat{f}(l_{\min}-1+g(1)), \\ f'(l_{\min}+1) &= \hat{f}(l_{\min}-1+g(2)), \\ &\vdots \\ f'(l_{\max})) &= \hat{f}(l_{\min}-1+g(l_{\max}-l_{\min}+1)). \end{aligned}$$

Now, suppose that $1 \leq i < j \leq l-1$. We aim to show that $b_{f'(i)f'(j)} \leq b_{f'(j)f'(i)}$.

First, assume that $i < l_{\min}$. Then, $f'(i) = \hat{f}(i) = f(i')$ where $i' = i$ if $i < k$ and $i' = i+1$ if $i \geq k$. If $i \geq k$, then $l_{\min} \geq k$ and so $l_{\min} = L_{\min} - 1$, implying that $i' < L_{\min}$. If $i < k$, then $i < l_{\min} \leq L_{\min}$. In either case $f'(i) = f(i')$ where $i' < L_{\min}$. We further note that the construction of $f'$ implies that $f'(j) = f(j')$ where $j' > i'$. Thus,

$$b_{f'(i)f'(j)} = b_{f(i')f(j')} \leq b_{f(j')f(i')} = b_{f'(j)f'(i)}.$$

Now, since $(i', j') \notin L$, we have either

$$b_{f(i')f(j')} \neq 0 \quad \text{or} \quad b_{f(i')f(k)}b_{f(k)f(j')} = 0.$$

As in the proof of case 1, if the first possibility holds we have $\pi_{f(i')} \geq \pi_{f(j')}$, for any positive diagonal $\Pi$ which symmetrises $A$, and if the second possibility holds we have

$$b_{f(i')f(j')} = b'_{f(i')f(j')} \quad \text{and} \quad b_{f(j')f(i')} = b'_{f(j')f(i')}.$$

Both possibilities imply that

$$b'_{f'(i)f'(j)} = b'_{f(i')f(j')} \leq b'_{f(i')f(j')} = b'_{f'(i)f(j)}.$$

The case $j > l_{\max}$ is very similar to that of $i < l_{\min}$. This assumption implies, as before, that $f'(j) = f(j')$ and $f'(i) = f(i')$ where $i' < j'$ and $(i', j') \notin L$. Thus, in this case we again have

$$b'_{f'(i)f'(j)} = b'_{f(i')f(j')} \leq b'_{f(i')f(j')} = b'_{f'(i)f(j)}.$$

So, we simply need to consider the case that $l_{\min} \leq i < j \leq l_{\max}$. The sequence $(g(1), \ldots, g(l_{\max} - l_{\min} + 1))$ is obtained by the reorder algorithm with input

$$\hat{B} = B'(\hat{f}(l_{\min}), \ldots, \hat{f}(l_{\max})).$$

Thus, for $i < j$, $\hat{b}_{g(i)g(j)} \leq \hat{b}_{g(j)g(i)}$. Then, we note that the $i'j'$th entry of $\hat{B}$ is the $\hat{f}(l_{\min} - 1 + i')\hat{f}(l_{\min} - 1 + j')$th entry of $B'$. As well, if $l_{\min} \leq i' \leq l_{\max}$,

$$f'(i') = \hat{f}(l_{\min} - 1 + g(i' - l_{\min} + 1)).$$

So, let $l_{\min} \leq i < j \leq l_{\max}$, let $i' = i - l_{\min} + 1$ and let $j' = j - l_{\min} + 1$. Then,

$$
\begin{aligned}
b'_{f'(i)f'(j)} &= b'_{\hat{f}(l_{\min}-1+g(i-l_{\min}+1))\hat{f}(l_{\min}-1+g(j-l_{\min}+1))} \\
&= b'_{\hat{f}(l_{\min}-1+g(i'))\hat{f}(l_{\min}-1+g(j'))} \\
&= \hat{b}_{g(i')g(j')} \\
&\leq \hat{b}_{g(j')g(j')} \\
&= b'_{\hat{f}(l_{\min}-1+g(j'))\hat{f}(l_{\min}-1+g(j'))} \\
&= b'_{\hat{f}(l_{\min}-1+g(j-l_{\min}+1))\hat{f}(l_{\min}-1+g(i-l_{\min}+1))} \\
&= b_{f'(j)f'(i)}.
\end{aligned}
$$

∎

The procedure behind the lower-weighted algorithm is the following. Let $A$ be a nearly uncoupled stochastic matrix. Suppose that the algorithm has proceeded through some number of iterations of its internal **while** loop; let $B$, $G$, $m$, $f$ and $l$ be the current stored data and let $k$ be the index selected by the algorithm (supposing that the algorithm will proceed through at least one more iteration). We assume that $B$ is error-reducing; as well, $B(f(1), \ldots, f(l))$ is lower-weighted and $k$ is the largest index such that

1. $b_{f(k)f(k)} < \frac{(1-\epsilon)^2}{1+(m_{f(k)}-2)\epsilon}$, and
2. the state $f(k)$ can be associated with a state $j$ that has a higher relative frequency (in the associated Markov chain).

That is, the first condition leads us to suspect that $k$ is the largest index such that $f(k)$ is not the sole remaining member of an almost invariant aggregate. We insist upon the second condition, as well, because the property that if $i \preceq_G j$ then $\pi(i) \leq \pi(j)$ is one of the base assumptions used to obtain the

$$\frac{(1-\epsilon)^2}{1+(m_{f(k)}-2)\epsilon}$$

bound in Appendix B. Thus, we assume that the stochastic complement $B \setminus f(k)$ is error-reducing as well.

Within the lower-weighted algorithm, it is not necessary to identify the collection $L$ and then reorder the submatrix $B(f(l_{\min}), \ldots, f(l_{\max}))$. One could simply re-calculate $f := \text{reorder}(B)$ at every iteration. However, we have found that, in practise, this makes the algorithm much less efficient.

Suppose that the matrix $B$ is a lower-weighted reversible stochastic matrix on the ordered state space $\mathcal{S}$ and let $\hat{B} = B \setminus i'$ be a stochastic complement. Let

$$L = \{(i,j) : i \neq j, \quad b_{ij} = 0 \quad \text{and} \quad b_{ii'}b_{i'j} \neq 0\}.$$

As we saw in the above proposition, if $(i,j) \notin L$ and $i < j$, then $\hat{b}_{ij} \leq \hat{b}_{ji}$. Thus, only the submatrix that contains all of the $ij$th entries where $(i,j) \in L$ needs to be reordered.

Moreover, as the algorithm proceeds, the successive stochastic complements have

significantly fewer 0-entries (the collection $L$ becomes smaller with successive complements). For example, suppose that the reversible stochastic matrix

$$B = \begin{bmatrix} \tilde{B} & v \\ w^T & b \end{bmatrix}$$

has $x$ nonzero off-diagonal entries. Let $x_1$ be the number of nonzero off-diagonal entries in the matrix $\tilde{B}$ and let $x_2$ be the number of nonzero entries in the vector $v$. Since $B$ is reversible, the vectors $v$ and $w$ have identical zero-nonzero patterns; so, there are also $x_2$ nonzero entries in $w$ and we have $x = x_1 + 2x_2$. The number of nonzero off-diagonal entries in the matrix

$$\frac{1}{1-b} v w^T$$

is $x_2^2 - x_2$ (there is one nonzero entry for each pair of distinct $i$ and $j$ with $v_i, w_j \neq 0$). So, the number of nonzero off-diagonal entries in the stochastic complement

$$\tilde{B} + \frac{1}{1-b} v w^T$$

is bounded above by $x_1 + x_2^2 - x_2 = x + x_2(x_2 - 3)$. The number of nonzero off-diagonal entries of $B$ can grow quite rapidly as we implement successive stochastic complements. Thus, the sizes of the submatrices that actually need to be reordered at each iteration can shrink equally rapidly.

## 6.3 The Perron-ordered algorithm

In some applications, the stationary distribution of a given stochastic matrix $A$ may be known. For example, let $X$ be the random walk on the weighted graph $G$, where the weight of the edge $ij$ is the $ij$th entry of the matrix $W$. Then, the vector $W\mathbb{1}$ is a scalar multiple of the stationary distribution of the transition matrix of $X$. As well, if the transition matrix has been obtained via a Markov chain Monte Carlo method, the stationary distribution is known (see Appendix C for an example of Markov chain Monte Carlo).

We present a simpler variation of the lower-weighted algorithm which includes the stationary distribution as an input; the inputs for the Perron-ordered algorithm are a reversible stochastic Matrix $A$ on the state space $\mathcal{S}$, the stationary distribution $\pi$ of $A$ and a nonnegative value $\epsilon < 1$.

If the original matrix $A$ is reversible, after any number of iterations of the algorithm, the matrix $B(f(1), \ldots, f(n))$ is lower-weighted, as $f$ is obtained from the stationary distribution (see the proof of Proposition 6.2). Thus, as applied to reversible matrices, the Perron-ordered algorithm is simply the lower-weighted algorithm with the calls to the reorder algorithm removed.

It may seem that the lower-weighted algorithm, applied to a reversible matrix $A$, is superfluous – one could simply calculate the stationary distribution $\pi$ of $A$ and then apply the Perron-ordered algorithm. However, we do not recommend this

**Algorithm 9** The Perron-ordered algorithm

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that contains no arcs.
$m := \mathbb{1}_{\mathcal{S}}$
$n := |\mathcal{S}|$
Let $f := (f(1), \ldots, f(n))$ be a bijection $\{1, \ldots, n\} \mapsto \mathcal{S}$ such that for $1 \leq i < j \leq n$, $\pi_{f(i)} \geq \pi_{f(j)}$.
**while** $n \geq 2$ **do**
  **if** for $k = 1, \ldots, n$, $b_{f(k)f(k)} \geq \frac{(1-\epsilon)^2}{1+(m_{f(k)}-2)\epsilon}$ or $\text{Choose}(B, f(k)) = 0$ **then**
    Exit the **while** loop.
  **else**
    $k := \max_{1 \leq k' \leq n} \left\{ k' : b_{f(k')f(k')} < \frac{(1-\epsilon)^2}{1+(m_{f(k')}-2)\epsilon} \text{ and } \text{Choose}(B, f(k')) \neq 0 \right\}$
    $j := \text{Choose}(B, f(k))$
    Add the directed arc $f(k) \to j$ to $G$.
    $m_j := m_j + m_{f(k)}$
    $B := B \setminus f(k)$
    $f := (f(1), \ldots, f(k-1), f(k+1), \ldots, f(n))$
    $n := n - 1$
  **end if**
**end while**
**return** $G$

approach. The stationary distribution $\pi$ is the unique solution to the eigenvalue problem $v^T A = v^T$. If the matrix $A$ is nearly uncoupled, it has a cluster of eigenvalues very near to 1, in which case the eigenvalue problem is referred to as badly conditioned. The output to a badly conditioned problem is very sensitive to measurement and round-off error, and, in principle, may be unreliable. For example, see [19] for an in depth discussion concerning the convergence of iterative techniques applied to nearly uncoupled Markov chains.

Thus, any potential vector produced as a solution to the eigenproblem $v^T A = v^T$ is possibly inaccurate, and may not be a reliable input to the Perron-ordered algorithm. The lower-weighted algorithm attempts to remove states with lower relative frequencies first, without actually calculating these frequencies.

The following line of reasoning suggests that if $A$ is not reversible, then the Perron-ordered algorithm is still reliable. Let $A$ be a nearly uncoupled stochastic matrix with stationary distribution $\pi$ on the state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate; let $B = A(\mathcal{E})$ and let $u = \pi(\mathcal{E})$. Let $\gamma = (I - B)\mathbb{1}$, so that for each $i \in \mathcal{E}$,

$$\mathbb{P}_i [x_1 \notin \mathcal{E}] = \gamma_i.$$

Consider the Markov chain $X$ with initial distribution equal to $\pi$. Thus, for any $t \geq 0$ and $i \in \mathcal{S}$

$$\mathbb{P} [x_t = i] = (\pi^T A^t)_i = \pi_i.$$

So,

$$\frac{u^T\gamma}{u^T\mathbb{1}} = \sum_{i\in\mathcal{E}} \frac{\mathbb{P}[x_t = i]}{\mathbb{P}[x_t \in \mathcal{E}]}\mathbb{P}[x_{t+1} \notin \mathcal{E} : x_t = i] = \mathbb{P}[x_{t+1} \notin \mathcal{E} : x_t \in \mathcal{E}].$$

**Definition 6.9.** Let $B$ be a substochastic matrix and let $u$ be a positive vector such that $u^T B \leq u^T$. We define the *u-weighted error* of $B$ to be the value

$$\eta_u(B) = \frac{u^T\gamma_B}{u^T\mathbb{1}} = \frac{u^T(I-B)\mathbb{1}}{u^T\mathbb{1}}.$$

We note that for any substochastic matrix $B$, $\eta(B) = \eta_{\mathbb{1}}(B)$. As well, let $A$ be an irreducible stochastic matrix with state space $\mathcal{S}$ and stationary distribution $\pi$, let $\mathcal{E} \subseteq \mathcal{S}$, let $B = A(\mathcal{E})$ and let $\hat{\pi} = \pi(\mathcal{E})$. Then, $\eta_{\hat{\pi}}(B) = 1 - w_\pi(B)$, the $\pi$-coupling measure of $B$ (and $\mathcal{E}$).

**Lemma 6.10.** *Let $B$ be a substochastic matrix on the state space $\mathcal{S}$ and let $i \in \mathcal{S}$ be such that the stochastic complement $\hat{B} = B \setminus i$ exists. Let $u$ be a positive vector such that $u^T B \leq u$ and suppose further that the subvector $\hat{u} = u(\mathcal{S} \setminus i)$ satisfies $\hat{u}^T\hat{B} \leq \hat{u}^T$. Then,*

$$\eta_{\hat{u}}(\hat{B}) \leq \frac{\eta_u(B)}{1-u_i}.$$

**Proof** Without loss of generality, we assume that $u^T\mathbb{1} = 1$ (multiplying $u$ by a positive scalar leaves $\eta_u(B)$ fixed). Let $\gamma = (I-B)\mathbb{1}$ and $\hat{\gamma} = (I-\hat{B})\mathbb{1}$. So,

$$\eta_u(B) = u^T\gamma.$$

Now, since $u^T B \leq u$, we have

$$\sum_{j \in \mathcal{S}} u_j b_{ji} \leq u_i,$$

further implying that

$$\sum_{j \neq i} u_j b_{ji} \leq u_i(1 - b_{ii}).$$

From the definition of the stochastic complement, it is straightforward to show that for each $j \in \mathcal{S} \setminus i$,

$$\hat{\gamma}_j = \gamma_j + \frac{b_{ji}}{1 - b_{ii}} \gamma_i.$$

So, we have

$$\begin{aligned}
\hat{u}^T \hat{\gamma} &= \sum_{j \neq i} u_j \hat{\gamma}_j \\
&= \sum_{j \neq i} u_j (\gamma_j + \tfrac{b_{ji}}{1-b_{ii}} \gamma_i) \\
&= \left( \sum_{j \neq i} u_j \gamma_j \right) + \tfrac{\gamma_i}{1-b_{ii}} \left( \sum_{j \neq i} u_j b_{ji} \right) \\
&\leq \left( \sum_{j \neq i} u_j \gamma_j \right) + \tfrac{\gamma_i}{1-b_{ii}} u_i (1 - b_{ii}) \\
&= \sum_{j \in \mathcal{S}} u_j \gamma_j \\
&= \eta_u(B).
\end{aligned}$$

We note that since $u^T \mathbb{1} = 1$ and $\hat{u}$ is obtained by deleting the $i$th entry from $u$, we have $\hat{u}^T \mathbb{1} = 1 - u_i$. Thus,

$$\eta_{\hat{u}}(\hat{B}) = \frac{\hat{u}^T \hat{\gamma}}{\hat{u}^T \mathbb{1}} \leq \frac{\eta_u(B)}{1 - u_i}.$$

∎

Let $A$ be a nearly uncoupled stochastic matrix with stationary distribution $\pi$ and state space $\mathcal{S}$; let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate. Lemma 6.10 informs us that if we remove the state $i \in \mathcal{E}$ via a stochastic complement, the probability of exiting $\mathcal{E}$ has been scaled upwards by a factor of at most $1/(1-\pi_i)$. Thus, choosing $i$ such that $\pi_i$ is minimal produces the best bound on this error inflation.

When the Perron Ordered Algorithm is applied to a nonreversible matrix $A$, any appearances of the expression

$$\frac{(1-\epsilon)^2}{1+(m-2)\epsilon}$$

(within the pseudocode) should be replaced with $(1-\epsilon)^m$.

## 6.4 The minimum column algorithm

We present an algorithm similar in spirit to the previous versions, which is intended for use with nearly uncoupled matrices $A$ which are not reversible and for which the stationary distribution is unknown.

Let $B$ be a substochastic matrix on the state space $\mathcal{S}$ containing 2 or more states and let $i \in \mathcal{S}$ be such that $b_{ii} < 1$. We define the $i$th modified column sum of $B$ to be the number

$$c_B(i) = \frac{1}{1-b_{ii}} \sum_{j \neq i} b_{ji}.$$

That is, the $i$th modified column sum is the sum of the off-diagonal entries in the $i$th column divided by $1 - b_{ii}$.

**Proposition 6.11.** *Let $B$ be an irreducible substochastic matrix on the state space $\mathcal{S}$ containing two or more states; let $i \in \mathcal{S}$ be such that $c_B(i)$ is minimal among states in $\mathcal{S}$ and let $\hat{B} = B \setminus i$. Then,*

$$\eta(\hat{B}) \leq \eta(B).$$

**Proof** Let $m \geq 2$ be the order of $B$. Since $\mathbb{1}^T B \mathbb{1} \leq m$, there is at least one $i' \in \mathcal{S}$ such that the sum of the entries in the $i'$th column of $B$ is less than or equal to 1. Then, for such a state $i'$,

$$\sum_{j \in \mathcal{S}} b_{ji'} \leq 1 \text{ implies that } \sum_{j \neq i'} b_{ji'} \leq 1 - b_{i'i'},$$

further implying that $c_B(i') \leq 1$. Thus, since $c_B(i)$ is minimal, $c_B(i) \leq 1$. Express

$$B \cong \begin{bmatrix} \tilde{B} & v \\ w^T & b_{ii} \end{bmatrix}$$

where the final row and column correspond to $i$. So,

$$c_B(i) = \frac{\mathbb{1}^T v}{1 - b_{ii}} \leq 1.$$

The statement can then be shown in the exact same manner as in Proposition 6.5 – in the proof there, the fact that $\eta(B \setminus i) \leq \eta(B)$ was deduced solely from the fact that $\mathbb{1}^T v \leq 1 - b_{ii}$. ∎

**Algorithm 10** The minimum column algorithm

$B := A$
Let $G$ be the digraph on $\mathcal{S}$ that contains no arcs.
$m := \mathbb{1}_{\mathcal{S}}$
$K := \{i \in \mathcal{S} : b_{ii} < 1 - \epsilon\}$
$\mathcal{C} := \emptyset$
**while** $|K| \geq 1$ **do**
  Let $i \in K$ be such that $c_B(i) = \min_{i' \in K}\{c_B(i')\}$.
  $\mathcal{C} := \mathcal{C} \cup \{i\}$
  Let $j \in \mathcal{S} \setminus \mathcal{C}$ be such that $b_{ij} = \max_{j' \in \mathcal{S} \setminus \mathcal{C}}\{b_{ij'}\}$.
  $B := B \setminus i$
  Add the directed arc $i \to j$ to $G$.
  $m_j := m_j + m_i$
  $K := \{k \in K \setminus i : b_{kk} < (1 - \epsilon)^{m_k}\}$
**end while**
**return** $G$

**Proposition 6.12.** *Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and suppose that we have applied Algorithm 10 to $A$. Let $B$, $K$, $\mathcal{C}$, $m$ and $G$ be the stored data after any number of iterations of the **while** loop. Then,*

1. *$B = A \setminus \mathcal{C}$,*
2. *$G$ is acyclic, every member of $\mathcal{C}$ has out-degree 1 in $G$ and every member of $\mathcal{S} \setminus \mathcal{C}$ has out-degree 0 in $G$,*
3. *for each $i \in \mathcal{S} \setminus \mathcal{C}$, $m_i$ is the order of the weakly connected component of the $G$ which contains $i$, and*
4. *$K = \{i \in \mathcal{S} \setminus \mathcal{C} : b_{ii} < (1 - \epsilon)^{m_k}\}$.*

The first three statements in Proposition 6.12 are shown as in Proposition 5.1 and Lemma 5.3. The fourth can be shown via a proof by induction.

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$ and let $0 \leq \epsilon < 1$. Suppose that we have applied Algorithm 10 to $A$ and let $B$, $K$, $\mathcal{C}$, $m$ and $G$ be the stored data after some number of iterations of the **while** loop; suppose further that $B$ is error reducing. In order to be sure that the next stochastic complement formed, $B \setminus i$, is error-reducing, we need to ensure that $c_{B'}(i)$ is minimal, for some unknown principal submatrix $B'$ of $B$. Since this matrix is unknown, we instead minimise $c_B(i)$, since $c_{B'}(i) \leq c_B(i)$ whenever $B'$ is a principal submatrix of $B$.

## 6.5 An algorithm for identifying near transient states

Let $A$ be a nearly uncoupled reversible stochastic matrix and let the digraph $G$ be formed by an application of one of our decoupling algorithms (Algorithms 3, 4, 8, 9 and 10). Let $V_1, \ldots, V_m$ be the vertex sets of the weakly connected components of $G$.

Recall that an $\epsilon$-uncoupling of $A$ is a partition $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$, where $\mathcal{E}_0$ is possibly empty and, for $k \neq 0$, each $\mathcal{E}_k$ is an almost invariant aggregate with respect to $\epsilon$.

In this section we present a method for constructing a potential $\epsilon$-uncoupling out

of the partition $(V_1, \ldots, V_m)$.

**Definition 6.13.** Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$ be nonempty. The *stochastic restriction* of $\mathcal{E}$ is the stochastic matrix $R(\mathcal{E})$ on the state space $\mathcal{E}$ defined via

$$r_{ij} = \begin{cases} a_{ij} & \text{if } \ i \neq j \\ 1 - \sum\limits_{k \in \mathcal{E} \setminus i} a_{ik} & \text{if } \ i = j. \end{cases}$$

The stochastic restriction is easily seen to be a stochastic matrix. First, its off-diagonal entries are nonnegative. Then, we note that its diagonal entries are nonnegative as well by observing that for each $i \in \mathcal{E}$

$$\sum_{k \in \mathcal{E} \setminus i} a_{ik} \leq \sum_{k \in \mathcal{S}} a_{ik} = 1.$$

Then, for each $i \in \mathcal{E}$,

$$\sum_{k \in \mathcal{E}} r_{ik} = r_{ii} + \sum_{k \in \mathcal{E} \setminus i} r_{ik} = 1 - \sum_{k \in \mathcal{E} \setminus i} a_{ik} + \sum_{k \in \mathcal{E} \setminus i} a_{ik} = 1.$$

So, the sum of the entries in each row of $R(\mathcal{E})$ is 1.

The stochastic restriction of a subspace $\mathcal{E}$ models the following Markov chain. We observe the Markov chain associated with the original stochastic matrix $A$ subject to the constraint $x_0 \in \mathcal{E}$. We add the further constraint that the Markov chain is "not allowed" to exit $\mathcal{E}$. We can imagine that every time a transition $i \to j$ where $i \in \mathcal{E}$ and $j \notin \mathcal{E}$ might occur, we replace this with the transition $i \to i$.

**Proposition 6.14.** *Let $A$ be an irreducible reversible stochastic matrix on the state space $\mathcal{S}$; let $\mathcal{E} \subseteq \mathcal{S}$ be such that $B = A(\mathcal{E})$ is irreducible and let $\hat{B} = R(\mathcal{E})$ be the stochastic restriction of $\mathcal{E}$. Let $\pi$ be the stationary distribution of $A$ and let $\hat{\pi}$ be the stationary distribution of $\hat{B}$. Then, $\hat{\pi}$ is a scalar multiple of the subvector $\pi(\mathcal{E})$ corresponding to $\mathcal{E}$.*

**Proof** Since $B$ is irreducible and the off-diagonal entries of $B$ and $\hat{B}$ are equal, the matrix $\hat{B}$ is an irreducible stochastic matrix and so has a unique stationary distribution $\hat{\pi}$. Moreover, the facts that $A$ is reversible and that $a_{ij} = \hat{a}_{ij}$ for all pairs of distinct $i$ and $j$ contained in $\mathcal{E}$ imply that

$$\pi_i \hat{b}_{ij} = \pi_j \hat{b}_{ji}$$

for all $i, j \in \mathcal{E}$. By proposition 2.12, the vector $\pi(\mathcal{E})$ is a scalar multiple of $\hat{\pi}$. Since $A$ is irreducible, every entry of $\pi$ is positive and the statement holds. ∎

Let $A$ be a nearly uncoupled reversible stochastic matrix on the state space $\mathcal{S}$ and let the digraph $G$ be formed by an application of one of our uncoupling algorithms. Let $V \subseteq \mathcal{S}$ be the vertex set of a weakly connected component of $G$. As in the discussion concerning Proposition 5.4, we suspect that $V$ consists of an almost invariant aggregate together with some number of near transient states.

Further, for each directed arc $i \to j$ present in $G$, there is a stochastic complement $\hat{A}$ of $A$ that has the $ij$th entry large. So, transitions within $V$ are very likely, whereas

transitions from $V$ to $\mathcal{S} \setminus V$ are less so. As well, since such a stochastic complement $\hat{A}$ is reversible, whenever $\hat{a}_{ij} \neq 0$ we also have $\hat{a}_{ji} \neq 0$. Thus, it seems reasonable to assume that $A(V)$ is irreducible.

Let $\pi$ be a stationary distribution of $A$ that is nonzero on the states contained in $V$ and label $V = \{i_1, \ldots, i_m\}$ so that

$$\pi_{i_1} \geq \pi_{i_2} \geq \ldots \geq \pi_{i_m}.$$

Near transient states are states that the associated Markov chain visits only rarely. For each $i, j \in V$, the ratio $\pi_j/\pi_i$ measures the relative frequency of visits to $i$ and $j$ – that is, the Markov chain visits state $j$ $\pi_j/\pi_i$ times as often as it visits state $i$. Thus, we will assume that either $V$ contains no near transient states or that for some $k$ with $2 \leq k \leq m$, the near transient states contained in $V$ are $i_k, i_{k+1}, \ldots, i_m$. That is, we assume that the near transient members of $V$ are those that have the smallest stationary weights.

Utilising this idea and Proposition 6.14, we propose the following algorithm for refining the output of our uncoupling algorithms.

We note that the following algorithm does not calculate the stationary distribution of the entire matrix $A$ – it calculates stationary distributions of stochastic restrictions of $A$ which we suspect are irreducible and "well-coupled". Therefore, even though the eigenvalue equation $v^T A = v^T$ is badly conditioned, we suspect that the eigenproblems we are solving, $w^T R(V) = w^T$, are well-conditioned.

Within the refining algorithm, we will use the 1-coupling measure previously introduced to evaluate the "strength" an aggregate. Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $B = A(\mathcal{C})$ be a principal submatrix of order $m \geq 1$; then,

$$w_1(B) = \frac{\mathbb{1}^T B \mathbb{1}}{m} = \frac{1}{m} \sum_{i,j \in \mathcal{C}} b_{ij}.$$

As usual, when $w_1(B)$ is close to 1, we suspect that $\mathcal{C}$ forms an almost invariant aggregate.

The inputs of the refining algorithm are a reversible stochastic matrix $A$ on the state space $\mathcal{S}$ and a partition $(V_1, \ldots, V_m)$ of $\mathcal{S}$. The output is a partition $(\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$ such that $\mathcal{E}_k \subseteq V_k$ (for $k \neq 0$),

$$\mathcal{E}_0 = \bigcup_{k=1}^{m} V_k \setminus \mathcal{E}_k$$

and each $w_1(A(\mathcal{E}_k))$ is maximal, within a certain class of submatrices of $A(V_k)$.

Let $A$ be a nearly uncoupled reversible stochastic matrix on the state space $\mathcal{S}$ and let the digraph $G$ be formed by an application of one of the uncoupling algorithms. Let $V \subseteq \mathcal{S}$ be the vertex set of a weakly connected component of $G$. The refining algorithm sorts the states in $V$ into descending under, under their weights in a stationary distribution of $A$. As we discussed above, the near transient states in $V$ should form the tail of this sequence. So, the refining algorithm simply calculates which leading portion of the sequence forms the strongest aggregate, under the 1-coupling measure.

We prefer the more conservative 1-coupling measure. However, the algorithm is

**Algorithm 11** The aggregate refining algorithm

**for** $t = 1, \ldots, s$ **do**
  $m := |V_t|$
  Calculate a stationary distribution $\hat{\pi}$ of the stochastic restriction $R(V_t)$.
  Let $f$ be a bijection $\{1, \ldots, m\} \mapsto V_t$ such that

$$\hat{\pi}_{f(1)} \geq \hat{\pi}_{f(2)} \geq \ldots \geq \hat{\pi}_{f(m)}.$$

  **for** $k = 1, \ldots, m$ **do**
    $w_k := \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{f(i)f(j)}$
  **end for**
  Let $k \in \{1, \ldots, m\}$ be such that $w_k = \max\{w_1, \ldots, w_m\}$.
  $\mathcal{E}_t := \{f(1), \ldots, f(k)\}$
**end for**
$\mathcal{E}_0 := \mathcal{S} \setminus \bigcup_{t=1}^{s} \mathcal{E}_t$
**return** $(\mathcal{E}_1, \ldots, \mathcal{E}_s, \mathcal{E}_0)$

already calculating subvectors of the stationary distribution; so, it is very straightforward to modify it to utilise the $\pi$-coupling measure instead. In this case, we simply replace the command

$$w_k := \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{f(i)f(k)}$$

with the command

$$w_k := \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \hat{\pi}_{f(i)} a_{f(i)f(k)}}{\sum_{i=1}^{k} \pi_{f(i)}}.$$

However, we note that the $\pi$-coupling measure already undervalues near transient states – thus the improvement to the strengths of the aggregates tends to be minimal when utilising the $\pi$-coupling measure.

# Chapter 7

# Conclusions and directions for future research

The stochastic complement based algorithms presented here are an efficient and effective tool for the construction of almost invariant aggregates of a given Markov chain. The three strengths of the approach are its efficiency, in terms of computation time required, its independence of spectral methods and the level of detail in its output. There are a number of unsolved problems regarding the application of these ideas; as well, we present sketches of potential future directions of this research.

## 7.1 Advantages of the approach

The speed at which the stochastic complement based algorithms operate is a definite point in their favour. Given a single stochastic matrix, even of relatively

large order, it is a straightforward computation task to compute many outputs (using alternate input values or more than one of our algorithms).

As we show in Appendix E, each of algorithms has a complexity bounded by $n^3$, where $n$ is the order of the input matrix. It is known that this the complexity of Gauss-Jordan Elimination (and many other important matrix-related algorithms).

For example, in Appendix C, we present a summary of the Lower Weighted Algorithm's performance when applied to a collection of randomly generated matrices. We generated 180 matrices of order 1000, and applied the Lower Weighted Algorithm to each matrix. This entire procedure took 87 minutes to execute, using MatLab 7 on a PC with a 2 GHz dual-core processor.

We suggest that the stochastic complement based algorithms' independence from spectral methods is another strength of the approach. Consider the following very simple example. Let

$$A_1 = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} 1-\epsilon^2 & \epsilon^2 \\ \epsilon & 1-\epsilon \end{bmatrix},$$

where $\epsilon < 1$ is some small positive constant. Now,

$$\|A_2 - A_1\|_\infty = \left\| \begin{bmatrix} \epsilon-\epsilon^2 & \epsilon^2-\epsilon \\ 0 & 0 \end{bmatrix} \right\|_\infty = 2(\epsilon-\epsilon^2).$$

The matrix $A_2$ can be viewed as a small perturbation of $A_1$. However, we find that this small perturbation of $A_1$ corresponds to a large perturbation of one of its right

eigenvectors. The Perron values of $A_1$ and $A_2$ are both $\rho = 1$ and the (right) Perron vector (of both matrices) is

$$v_\rho = \frac{1}{2}\begin{bmatrix} 1 \\ \\ 1 \end{bmatrix}.$$

The second right eigenpairs of $A_1$ and $A_2$ are given by $A_1 v_1 = (1 - 2\epsilon)v_1$ and $A_2 v_2 = (1 - \epsilon - \epsilon^2)v_2$, where

$$v_1 = \begin{bmatrix} 1 \\ \\ -1 \end{bmatrix} \text{ and } v_2 = \begin{bmatrix} \epsilon \\ \\ -1 \end{bmatrix}.$$

We have normalised all of the above vectors so that each has a $\infty$-norm of 1. We calculate

$$\|v_2 - v_1\|_\infty = \left\| \begin{bmatrix} \epsilon - 1 \\ \\ 0 \end{bmatrix} \right\|_\infty = 1 - \epsilon.$$

Even though the value of $\|A_2 - A_1\|_\infty$ is on the order of $\epsilon$, the value $\|v_2 - v_1\|_\infty$ is close to 1 (a significant difference for normalised vectors).

The collections $C_1 = \{v_\rho, v_1\}$ and $C_2 = \{v_\rho, v_2\}$ are the basis upon which the Perron cluster approach partitions the state spaces of $A_1$ and $A_2$ (if it were to be applied to these matrices). That is, a small perturbation of the matrix $A_1$ results in a large perturbation of the eigenvectors associated with the Perron cluster.

We produce this example to show that when a stochastic matrix is nearly uncoupled, its spectral properties (and thus, the information upon which spectral based

algorithms operate) can be wildly sensitive to perturbation. In the example above, the collections $C_1$ and $C_2$ induce identical (and correct) decompositions of the state space. However, for matrices that are of much larger order, and which possess less straightforward nearly uncoupled structure, it is unclear what effect such tiny perturbations may have on the vectors associated with the Perron cluster. We propose that an approach that does not rely on such sensitive structures is desirable.

The final point which we raise, in our approach's support, is the level of detail of its output. Other approaches (for example, the Perron cluster and SVD based algorithms) use a partitioning approach. One begins with the state space $\mathcal{S}$, and then partitions it into steadily smaller subsets until an $\epsilon$-uncoupling is achieved.

We use an aggregating approach – one begins with the collection of singleton sets, $\Psi = (\{i\})_{i \in \mathcal{S}}$, and then takes unions, forming larger and larger sets, until an uncoupling is constructed. The advantage of this method is that if we "save our work", it is straightforward to construct subaggregates of the produced collections – we simply use elements of the previously constructed partitions. Other uses for this hierarchical structure can be constructed – for example, the recursive subaggregating procedure we present in Appendix C.

## 7.2 Improvement of the bound in Appendix B

In Appendix A we show the following. Let $A$ be an irreducible reversible stochastic matrix and let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate, containing 2 or more states, with respect to $\epsilon$. Let $\Pi$ be a positive diagonal matrix such that $\Pi A$ is symmetric, let $i \in \mathcal{E}$ be such that

$$\pi_i = \max_{j \in \mathcal{E}} \{\pi_j\}$$

and let $\tilde{A} = A \setminus \{j \in \mathcal{E} : j \neq i\}$. Then,

$$\tilde{a}_{ii} \geq \frac{(1-\epsilon)^2}{1 + (|\mathcal{E}| - 2)\epsilon}.$$

This is our motivation for using

$$\tilde{a}_{ii} \geq \frac{(1-\epsilon)^2}{1 + (m_i - 2)\epsilon}$$

as the test for whether or not it is safe to remove state $i$ in the Modified Maximum Entry, Lower Weighted and Perron Ordered Algorithms. These algorithms have been specifically constructed so that they do not remove states with maximal $\Pi$-values.

Let $A$ be a stochastic matrix on the state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate of $m$ states with respect to $\epsilon$. In Appendix B we show that there is at least one index $i \in \mathcal{E}$ of such that $\tilde{A} = A \setminus \{j \in \mathcal{E} : j \neq i\}$ has

$$\tilde{a}_{ii} \geq (1-\epsilon)^m.$$

However, we have not yet identified necessary or sufficient conditions that identify such a state $i$. It is unknown, at this point, if the Minimum Column Algorithm, or any of our stochastic complement based algorithms, will fail to remove such states first.

We believe that the following conjecture holds. Let $B = B^{(0)}$ be an irreducible substochastic matrix of order $m \geq 2$ such that $B\mathbb{1} \geq (1-\epsilon)\mathbb{1}$. Let $B^{(1)}, \ldots, B^{(m-1)}$ be a sequence of stochastic complements such that $B^{(k+1)} = B^{(k)} \setminus i$ where the $i$th column sum of $B^{(k)}$ has minimal sum. Then, $B^{(m-1)} = [\alpha]$ where $\alpha \geq (1-\epsilon)^m$. Moreover, we suspect that equality occurs if and only if $B = (1-\epsilon)P$ where $P$ is a cyclic permutation matrix (as in Proposition B.12).

This conjecture is very simple to prove for $m = 2$. However, it seems to be challenging to show that it holds in general, or even for the $m = 3$ case. If this conjecture can be shown to be true, we would have a stronger basis for utilising the bound described in Appendix B.

## 7.3 Mean first passage times

Each of our stochastic complement based algorithms uses the following idea in its implementation.

Let $A$ be a stochastic matrix and let $\tilde{A} = A\backslash\mathcal{C}$ be a stochastic complement formed after some number of iterations of one of the complement based algorithms. Suppose

that state $i$ is the next state the algorithm will remove. Let state $j \notin \mathcal{C}$, distinct from $i$, be such that

$$\tilde{a}_{ij} = \max_{\substack{k \notin \mathcal{C} \\ \ni k \neq i}} \{\tilde{a}_{ik}\}.$$

Then, states $i$ and $j$ (and all other states already associated with $i$ and $j$) will be connected by the algorithm. That is, a directed arc $i \to j$, where $j$ satisfies the above equality, will be added to the output digraph.

An important open problem remaining is the following. If the Markov chain associated with $A$ is nearly uncoupled and $\tilde{a}_{ij}$ is as above, under what circumstances can we be sure that $i$ and $j$ belong to a minimal almost invariant aggregate? It is fairly straightforward to construct examples where $i$ and $j$ belong to distinct almost invariant aggregates. For example, in Appendix D we produce a characterisation of block homogeneous stochastic matrices which identifies exactly when this condition holds and when it fails.

We consider this problem, briefly. Let $X$ be a Markov chain on the state space $\mathcal{S}$ with transition matrix $A$. As before, the random variable

$$T_i = \inf\{t \geq 1 : x_t = i\}$$

(with the convention that $\inf \emptyset = \infty$) is the stopping time referred to as the first passage time into $i$. For each $i, j \in \mathcal{S}$, we refer to the value

$$t_{ij} = \mathbb{E}\left[T_j | x_0 = i\right]$$

as the *mean first passage time* from $i$ to $j$. That is, given $x_0 = i$, $t_{ij}$ is the expected value of the smallest positive $t$ with $x_t = j$. We note that $t_{ij} \geq 1$, and that it is entirely possible that $t_{ij} = \infty$.

Proposition 7.1 appears in [16, Lemma 2.2].

**Proposition 7.1.** *Let $X$ be a Markov chain on the state space $\mathcal{S}$ with transition matrix $A$ and suppose that $A$ is irreducible. Let $j \in \mathcal{S}$ and express*

$$A \cong \begin{bmatrix} a_{jj} & w^T \\ v & B \end{bmatrix}.$$

*For each $i \neq j$, $t_{ij}$ is equal to the $i$th entry of $(I - B)^{-1}\mathbb{1}$.*

**Proposition 7.2.** *Let $\epsilon < 1$ be positive and let $A$ be a stochastic matrix of the form*

$$A = \begin{bmatrix} B_1 & B_{12} \\ B_{21} & B_2 \end{bmatrix}$$

*where $B_1\mathbb{1} \geq (1 - \epsilon)\mathbb{1}$ and $B_2\mathbb{1} \geq (1 - \epsilon)\mathbb{1}$. Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be the collections of states corresponding to the block expression of $A$. Suppose that whenever $i$ and $j$ are contained in the same member of $\{\mathcal{E}_1, \mathcal{E}_2\}$ we have $t_{ij} < 1/\epsilon$. Then, each of $B_1$ and $B_2$ either has at least one off-diagonal entry strictly greater than $\epsilon$ or is $1 \times 1$.*

**Proof** We proceed by contradiction. Suppose that $B_1$ is of order 2 or greater and that every off-diagonal entry of $B_1$ is less than or equal to $\epsilon$. Express

$$A = \begin{bmatrix} a_{11} & w^T \\ v & A' \end{bmatrix}.$$

For each $i \in \mathcal{E}_1 \setminus 1$, we have $a_{i1} \leq \epsilon$, by assumption (for any $i \in \mathcal{E}_1$, the entry $a_{i1}$ is contained in the block $B_1$). For each $j \in \mathcal{E}_2$, the entry $a_{j1}$ is contained in the block $B_{21}$, and so is less than or equal to $\epsilon$. Thus, $v \leq \epsilon \mathbb{1}$, further implying that $A'\mathbb{1} \geq (1-\epsilon)\mathbb{1}$. This implies that

$$(I - A')\mathbb{1} \leq \epsilon \mathbb{1},$$

and so we see that

$$\frac{1}{\epsilon}\mathbb{1} \leq (I - A')^{-1}\mathbb{1}.$$

By Proposition 7.1, we must have $t_{i1} \geq 1/\epsilon$ for all $i \neq 1$. Thus, if $t_{ij} < 1/\epsilon$ for all distinct pairs $i, j \in \mathcal{E}_1$, either there are no such pairs ($B_1$ is $1 \times 1$) or there is at least one off-diagonal entry of $B_1$ strictly greater than $\epsilon$. The same is true of $B_2$, via similarity of the argument. ∎

Using Proposition 7.2, we can prove the following. Let $A$ be a nearly uncoupled stochastic matrix and suppose that there is an $\epsilon$-uncoupling $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m, \mathcal{E}_0)$ such that for each $k \neq 0$, if $i, j \in \mathcal{E}_k$, then $t_{ij} < 1/\epsilon$. Then, the very first directed arc $i \to j$ with $i \notin \mathcal{E}_0$ added to the output digraph by the Maximum Entry Algorithm (with input $\epsilon$) has $i, j \in \mathcal{E}_k$ for some $k \neq 0$.

As we stated above, a problem remaining in the study of these stochastic complement based algorithms is to find conditions that guarantee that further iterations of the Maximum Entry Algorithm, or any iterations of our other algorithms, are

also correct. We hypothesise that some set of assumptions concerning the mean first transition times and the sizes of the minimal aggregates will be sufficient. However, it seems to be a very challenging problem to find such conditions that still retain a useful degree of generality.

## 7.4 Recommender systems

An interesting application of matrix theory is to so-called *recommender systems*. Put simply, a recommender system is an algorithm or method which recommends entries of a database to its users. Generally, these recommendations are based on the users' past histories of interactions with the database – a recommender system attempts to guess which entries are appropriate or desirable to each of its users. A survey of such systems is found in [1].

As an example, we provide a sketch of the system used by Amazon.com to make product recommendations to users browsing its online store; see [17] for an introduction to the company's algorithm. Let $P$ be the collection of products offered; for each $x \in P$, let $n(x)$ be the number of customers who have purchased product $x$ and for each pair of distinct $x, y \in P$, let $n(x, y)$ be the number of customers who have purchased both products. For any pair of distinct products $x, y \in P$, the similarity between $x$ and $y$ is the value

$$\text{sim}(x, y) = \frac{n(x, y)}{\sqrt{n(x)n(y)}}.$$

For any two products $x, y \in P$,

1. $0 \leq \text{sim}(x, y) \leq 1$,

2. $\text{sim}(x, y) = 1$ if and only if any customer which purchased one of $x$ or $y$ purchased both of $x$ and $y$, and

3. $\text{sim}(x, y) = 0$ if and only if no customer purchased both $x$ and $y$.

The similarity between two products is an attempt to measure how likely a customer who has purchased one is to have purchased both.

Now, suppose that a customer is browsing Amazon.com; let $V$ be the collection of products which the customer has purchased or viewed (including the product the customer is currently viewing). The Amazon.com system simply recommends products $y \notin V$ such that for one or more $x \in V$, $\text{sim}(x, y)$ is relatively high.

The stochastic complement can be used to enhance such a system in two ways. We illustrate both using Amazon.com's recommender system.

Let $G$ be the weighted graph with vertices equal to $P$ (as above) where the weight of edge $xy$ is equal to $\text{sim}(x, y)$. (We may or may not include loops at the vertices – for the purposes of this sketch, this choice is not relevant.) Let $A$ be the transition matrix of the random walk on $G$. Then, given $V$ as above, the Amazon.com is recommending products $y \notin V$ such that for one or more $x \in V$, $a_{xy}$ is close to

$$\max_{x' \in V, y' \notin V} \{a_{x'y'}\}.$$

A first use for the stochastic complement based approach is to run one of our algorithms in order to detect almost invariant subsets of $P$ (with respect to $A$). It seems likely that such collections would occur – the Amazon.com graph of products is known to be large and sparse. Suppose that a customer is browsing the online store and let $V$ be as above. Then, if there is an almost invariant aggregate $\mathcal{E}$ such that $|V \cap \mathcal{E}|/|V|$ is significant, it would seem prudent to recommend members of $\mathcal{E}$ to the customer, especially products $y \in \mathcal{E}$ that have been purchased by large numbers of customers.

A second way in which the stochastic complement may supplement such a recommender system is the following. Suppose that a customer is browsing the online store and let $V$ be as above. Let $V' \subseteq V$ be such that $V \setminus V' = \{z\}$. (That is, let $V'$ contain every member of $V$ except for one, which we label $z$.) Let $A$ be as above and let $\tilde{A} = A \setminus V'$. Rather than selecting products $y \notin V$ where $a_{xy}$ is close to

$$\max_{x' \in V, y' \notin V} \{a_{x'y'}\}$$

(for some $x \in V$), we suggest that it could useful to recommend products $y \notin V$ such that $\tilde{a}_{zy}$ is relatively close to

$$\max_{y' \notin V} \{\tilde{a}_{zy'}\}.$$

This allows a customer's history of purchases and views to be more fully incorporated into the recommendations.

It may be that there are products $y \notin V$ such that $\text{sim}(x, y)$ is small for all $x \in V$, but the value

$$\sum_{x \in V} \text{sim}(x, y)$$

is relatively large. Such products seems like good candidates to be recommended to the customer. The stochastic complement tends to preserve such structures – such recommendations may be made more likely than with Amazon.com's method.

Candidates for the product $z \in V$ (described above) are the product the customer is currently viewing or a product included in the customer's most recent purchase. Such a system attempts to anticipate the customer's next purchase, incorporating their full purchasing and viewing history.

We note that the second suggestion partially implements the first, at least implicitly. If there is an almost invariant aggregate $\mathcal{E}$ from which the customer is making large numbers of purchases, the values $\tilde{a}_{zy}$, where $y \in \mathcal{E}$ and $z$ is as above seem likely to become large.

In future works we aim to flesh out such ideas more fully and explore the properties of specific implementations of the stochastic complement to recommender systems (and other data mining concepts).

# Appendix A

# A lower bound concerning stochastic complements of reversible Markov chains

We construct a lower bound on a specific term relating to stochastic complements of reversible substochastic matrices.

## A.1 Definitions and problem statement

**Definition A.1.** Let $B$ be a properly substochastic matrix and let $\mathcal{C}$ be the associated state space. If the order of $B$ is 1, that is, if $\mathcal{C} = \{i\}$ and $B = [b_{ii}]$, we define $\alpha_B(i) = b_{ii}$. If $\mathcal{C}$ contains two or more states, then for each $i \in \mathcal{C}$ we express

$$B \cong \begin{bmatrix} b_{ii} & v^T \\ w & A \end{bmatrix},$$

and define

$$\alpha_B(i) = b_{ii} + v^T(I - A)^{-1}w.$$

An alternate way to define $\alpha_B(i)$ is the following. Let $B$ and $\mathcal{C}$ be as in Definition A.1 and let $i \in \mathcal{C}$. Let $\hat{B}$ be the stochastic complement that removes every state aside from $i$; that is, $\hat{B} = B \setminus \{j : j \neq i\}$. Then, $\hat{B}$ is the $1 \times 1$ substochastic matrix

$$\hat{B} = [\alpha_B(i)].$$

Let $X$ be a Markov chain on the state space $\mathcal{S}$. Recall that, for $\mathcal{C} \subseteq \mathcal{S}$, we define

$$E_{\mathcal{C}} = \inf_{t \geq 1} \{t : x_t \notin \mathcal{C}\}.$$

If $x_0 \in \mathcal{C}$, we refer to $t = E_{\mathcal{C}}$ as the first exit time out of $\mathcal{C}$ and we say that the Markov chain exits $\mathcal{C}$ at time $t$. As well, for each $i \in \mathcal{S}$,

$$T_i = \inf_{t \geq 1} \{t : x_t = i\}$$

is the first passage time into $i$.

**Proposition A.2.** *Let $X$ be an irreducible Markov chain with state space $\mathcal{S}$ and transition matrix $A$. Let $\mathcal{C} \subseteq \mathcal{S}$ and let $B = A(\mathcal{C})$. For each $i \in \mathcal{C}$, $\alpha_B(i)$ is the probability of transitioning from $i$ to $i$, in one or more steps, without first exiting $\mathcal{C}$. That is,*

$$\alpha_B(i) = \mathbb{P}_i [T_i < E_{\mathcal{C}}].$$

**Proof** First we consider the case that $\mathcal{C} = \{i\}$. In this case, we simply have $\alpha_B(i) = b_{ii} = a_{ii}$. Clearly, $\alpha_B(i)$ is the probability of transitioning from $i$ to $i$ without first visiting any other state.

So, assume that $|\mathcal{C}| = m \geq 2$. Without loss of generality, we assume that $i = 1$ and that $\mathcal{C} = \{1, \ldots, m\}$ where $2 \leq m \leq |\mathcal{S}|$. Let $\mathcal{C}' = \mathcal{C} \setminus 1 = \{2, \ldots, m\}$. Express

$$A = \begin{bmatrix} a_{11} & v_1^T & v_2^T \\ w_1 & B_1 & B_{12} \\ w_2 & B_{21} & B_2 \end{bmatrix},$$

where the second row and column of blocks corresponds to states $\mathcal{C}'$ and the third corresponds to states $\mathcal{S} \setminus \mathcal{C} = \{m+1, \ldots, n\}$ (where $n$ is the order of $A$). The final column and row of blocks may be null; *i.e.* we may have $\mathcal{C} = \mathcal{S}$. However, our calculations will not include any of these terms, and so the presence or absence of these blocks is irrelevant.

Now,

$$B = \begin{bmatrix} a_{11} & v_1^T \\ w_1 & B_1 \end{bmatrix}$$

and $\alpha_B(1) = a_{11} + v_1^T(I - B_1)^{-1}w_1$. We have

$$A \setminus \mathcal{C}' = \begin{bmatrix} a_{11} & v_2^T \\ w_2 & B_2 \end{bmatrix} + \begin{bmatrix} v_1^T \\ B_{21} \end{bmatrix} (I - B_1)^{-1} \begin{bmatrix} w_1 & B_{12} \end{bmatrix} = \begin{bmatrix} \alpha_B(1) & * \\ * & * \end{bmatrix}.$$

(Only the $(1,1)$th entry is relevant to our discussion.) By Proposition 4.3, $\alpha_B(1)$ is the probability that, given $x_0 = 1$, there is some positive integer $t' \geq 1$ such that

$$\{x_t : 1 \leq t \leq t' - 1\} \subseteq \mathcal{C}'$$

and $x_{t'} = 1$. We show that this occurs if and only if $T_1 < E_{\mathcal{C}}$.

Suppose that $x_0 = 1$,

$$\{x_t : 1 \leq t \leq t' - 1\} \subseteq \mathcal{C}'$$

and $x_{t'} = 1$. Then, $T_1 = t'$, since $x_{t'} = 1$ and if $1 \leq t \leq t' - 1$ then $x_t \neq 1$. As well, $E_{\mathcal{C}} > t'$, since $x_1, \ldots, x_{t'} \in \mathcal{C}$. Thus, if the positive integer $t'$ satisfies these conditions, then $t' = T_1 < E_{\mathcal{C}}$.

Now, suppose that $T_1 < E_{\mathcal{C}}$. This implies that $T_1 \neq \infty$, as $E_{\mathcal{C}} \leq \infty$. Let $t' = T_1$, so that $x_{t'} = 1$ and $x_t \neq 1$ whenever $1 \leq t \leq t' - 1$. As well, if $1 \leq t \leq t' - 1$ then $1 \leq t < E_{\mathcal{C}}$, implying that $x_t \in \mathcal{C}$. Thus, if $x_0 = 1$ and $T_1 < E_{\mathcal{C}}$ then there is $t' \geq 1$ such that

$$\{x_t : 1 \leq t \leq t' - 1\} \subseteq \mathcal{C}'$$

and $x_{t'} = 1$. ∎

We recall that we refer to the substochastic matrix $B$ as reversible if there is a positive diagonal matrix $\Pi$ where $\Pi B$ is symmetric. As before, we use $\pi_i$ to represent the $i$th diagonal entry of the diagonal matrix $\Pi$.

**Lemma A.3.** *Let $B$ be an irreducible reversible substochastic matrix. Then, the positive diagonal matrices that symmetrise $B$ via left-multiplication are uniquely defined, up to multiplication by a positive constant.*

**Proof** We aim to show that if $\Pi$ and $\Pi'$ are positive diagonal matrices such that $\Pi B$ and $\Pi' B$ are symmetric, then $\Pi' = p\Pi$, for some positive scalar $p$. If $B$ is a $1 \times 1$ matrix, this is trivial; so, we assume that the order of $B$ is 2 or more. Thus, the digraph $G$ associated with $B$ is strongly connected.

Let $\Pi$ be a positive diagonal matrix such that $\Pi B$ is symmetric. Let

$$Q = \frac{1}{\pi_1}\Pi;$$

so, $QB$ is symmetric, as well. Thus, for all $i$ and $j$,

$$q_i b_{ij} = q_j b_{ji}.$$

This implies that if $b_{ij} \neq 0$, then $b_{ji} \neq 0$ and the ratio $q_i/q_j = b_{ji}/b_{ij}$ is uniquely determined by $B$. Let $i \neq j$ be any two distinct indices of $B$. Since the digraph of $B$ is strongly connected, there is directed walk from $i$ to $j$,

$$i = i_0 \to i_1 \to \cdots \to i_k = j,$$

present in $G$. So, if $0 \leq s \leq k-1$, then $b_{i_s i_{s+1}} \neq 0$. The above observation implies that each ratio $q_{i_s}/q_{i_{s+1}}$ is a positive scalar which is uniquely determined by $B$. This in turn implies that the ratio $q_i/q_j$ is positive and is uniquely determined by the

substochastic matrix $B$. Since $q_1 = 1$, each of the numbers $q_i = q_i/q_1$ is uniquely determined by $B$. Thus, if $\Pi'$ is a second positive diagonal matrix such that $\Pi' B$ is symmetric, then

$$\frac{1}{\pi_1}\Pi = Q = \frac{1}{\pi_1'}\Pi'.$$

∎

We use this lemma to uniquely identify the positive diagonal matrices associated with reversible substochastic matrices. Let $B$ be an irreducible reversible substochastic matrix with states $\mathcal{C}$; we define $\Pi = \Pi_B$ to be the unique positive diagonal matrix such that $\Pi B$ is symmetric and the largest diagonal entry of $\Pi$ is 1:

$$\max_{i \in \mathcal{C}} \{\pi_i\} = 1.$$

**Definition A.4.** Let $n \geq 1$ be a positive integer and $\epsilon < 1$ be a positive real number. We define $\mathcal{B}(n, \epsilon) = \{B\}$ to be the collection of $n \times n$ substochastic matrices $B$ such that

1. $B$ is irreducible and reversible, and
2. $\gamma_B = (I - B)\mathbb{1} \leq \epsilon \mathbb{1}$.

We note that for all $B \in \mathcal{B}(n, \epsilon)$, $B\mathbb{1} \geq (1 - \epsilon)\mathbb{1}$.

**Definition A.5.** Let $n \geq 1$ and let $\epsilon < 1$ be a positive real number. Let $B \in \mathcal{B}(n, \epsilon)$ and let $\Pi = \Pi_B$. We define $\alpha(B)$ to be the minimum value of $\alpha_B(i)$ subject to $\pi_i = 1$:

$$\alpha(B) = \min_{i \ni \pi_i = 1} \{\alpha_B(i)\}.$$

We note that if $B \in \mathcal{B}(n, \epsilon)$ and $\Pi = \Pi_B$, then for every index $i$ of $B$, either $\pi_i < 1$ or $\alpha_B(i) \geq \alpha(B)$.

The problem we solve is the following. Given positive integer $n \geq 2$ and positive real number $\epsilon < 1$, we calculate the number

$$\alpha(\mathcal{B}) = \inf_{B \in \mathcal{B}} \{\alpha(B)\}$$

and characterise those reversible substochastic matrices $B \in \mathcal{B}$ that have $\alpha(B) = \alpha(\mathcal{B})$.

## A.2 Preliminaries

**Lemma A.6.** *Let $B \in \mathcal{B}(n, \epsilon)$ where $n \geq 2$. Then, we can express*

$$B \cong \begin{bmatrix} a & v^T \\ w & A \end{bmatrix}$$

*where, in addition to the fact that $B$ is irreducible and substochastic,*

1. $a + v^T \mathbb{1} \geq 1 - \epsilon$,

2. $A\mathbb{1} + w \geq (1 - \epsilon)\mathbb{1}$,

3. $\alpha(B) = a + v^T (I - B)^{-1} w$, *and*

*4. there is a positive diagonal matrix $Q$, such that $Q \leq I$, $QA = A^T Q$ and $Qw = v$.*

**Proof** Let $\Pi = \Pi_B$. Since

$$\alpha(B) = \min_{i \ni \pi_i = 1} \{\alpha_B(i)\},$$

there is an index $i$ such that $\pi_i = 1$ and $\alpha(B) = \alpha_B(i)$. Express

$$B \cong \begin{bmatrix} b_{ii} & v^T \\ w & A \end{bmatrix}$$

where the first row and column corresponds to such a state $i$ and the principal submatrix $A$ corresponds to the remainder of the state space. The first two claims are direct consequences of the fact that

$$\gamma_B = (I - B)\mathbb{1} \leq \epsilon \mathbb{1}.$$

The third claim is simply a restatement of the fact that $\alpha_B(i) = \alpha(B)$.

Finally, since $\pi_i = 1$ and $\pi_j \leq 1$ for all $j \in \mathcal{C}$, we have

$$\Pi_B \cong \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix},$$

where $Q \leq I$, via the same correspondence as $B$. The fourth claim is a consequence of the fact that $\Pi_B B = B^T \Pi_B$. ∎

**Lemma A.7.** *Let $X$, $Y$ and $Z$ be nonnegative square matrices of order $m \geq 1$ such that*

*1. $Z$ is irreducible,*

*2. $X$ and $Y$ are positive diagonal matrices,*

*3. $X \leq Y$, with strict inequality in at least one diagonal entry, and*

*4. $Z\mathbb{1} \leq X\mathbb{1}$, with strict inequality in at least one position.*

*Then, the matrices $(X - Z)^{-1}$ and $(Y - Z)^{-1}$ are defined and satisfy*

$$0 < (Y - Z)^{-1} < (X - Z)^{-1}$$

*(entrywise).*

**Proof** A matrix is irreducible if its associated digraph is strongly connected or if it is the 0-matrix of order 1. If $Z = [0]$, we then have $X = [x]$ and $Y = [y]$ where $0 < x < y$. Then, $(X - Z)^{-1} = [1/x]$ and $(Y - Z)^{-1} = [1/y]$, where $0 < 1/y < 1/x$. So, we assume that the digraph $G$ associated with $Z$ is strongly connected.

Since $Z$ is irreducible and $Z\mathbb{1} \leq X\mathbb{1} \leq Y\mathbb{1}$, with each inequality strict in at least one position, $X^{-1}Z$ and $Y^{-1}Z$ are irreducible properly substochastic matrices. Thus,

$$(I - X^{-1}Z)^{-1} = \sum_{s \geq 0} (X^{-1}Z)^s \text{ and } (I - Y^{-1}Z)^{-1} = \sum_{s \geq 0} (Y^{-1}Z)^s$$

exist and are entrywise nonnegative (Lemma 2.8). So,

$$(X - Z)^{-1} = (I - X^{-1}Z)^{-1}X^{-1} = \sum_{s \geq 0} (X^{-1}Z)^s X^{-1}$$

$$\text{and } (Y-Z)^{-1} = (I - Y^{-1}Z)^{-1}Y^{-1} = \sum_{s\geq 0}(Y^{-1}Z)^s Y^{-1}$$

are entrywise nonnegative. Let $G$ be the directed graph induced by $Z$. We will use the abbreviations $x_i = x_{ii}$ and $y_i = y_{ii}$ to refer to the diagonal entries of $X$ and $Y$. For each directed walk

$$\omega = i_0 \to i_1 \to \ldots \to i_l$$

of length $l \geq 1$ in $G$ we define

$$x(\omega) = \frac{z_{i_0 i_1} z_{i_1 i_2} \cdots z_{i_{l-1} i_l}}{x_{i_0} x_{i_1} \cdots x_{i_l}} \text{ and } y(\omega) = \frac{z_{i_0 i_1} z_{i_1 i_2} \cdots z_{i_{l-1} i_l}}{y_{i_0} y_{i_1} \cdots y_{i_l}}.$$

For the directed walk $\omega = i$ of length 0 (the walk consisting of $i$ and no directed arcs), we define

$$x(\omega) = \frac{1}{x_i} \text{ and } y(\omega) = \frac{1}{y_i}.$$

By our above formulation, we have

$$\left[(X-Z)^{-1}\right]_{ij} = \sum_{\omega: i \rightsquigarrow j} x(\omega) \text{ and } \left[(Y-Z)^{-1}\right]_{ij} = \sum_{\omega: i \rightsquigarrow j} y(\omega).$$

We note that, since $X \leq Y$, $0 < y(\omega) \leq x(\omega)$ for all directed walks $\omega$ in $G$. Let $k$ be such that $x_k < y_k$; let $i$ and $j$ be any two indices (possibly identical to each other and/or to $k$). Since $Z$ is irreducible, there is a directed walk $\omega'$ in $G$ from $i$ to $j$ that

visits $k$ at least once. Such a walk has $0 < y(\omega') < x(\omega')$. We see that for any $i$ and $j$,

$$0 < \sum_{\omega: i \rightsquigarrow j} y(\omega) < \sum_{\omega: i \rightsquigarrow j} x(\omega).$$

Therefore,

$$0 < \left[(Y - Z)^{-1}\right]_{ij} < \left[(X - Z)^{-1}\right]_{ij}$$

for any $i$ and $j$. ∎

**Lemma A.8.** *Let $z \mapsto A(z)$ be a matrix-valued function $\mathbb{R} \mapsto \mathbb{R}_{n \times n}$ where each $a_{ij}(z)$ is a differentiable function of $z$. At points $z_0$ where $A(z_0)$ is nonsingular, we denote the inverse $(A(z_0))^{-1} = A^{-1}(z_0)$. Let*

$$\frac{d}{dz} A(z) = \left[\frac{d}{dz} a_{ij}(z)\right].$$

*If $A(z_0)$ is nonsingular, then*

$$\frac{d}{dz} A^{-1}(z)\bigg|_{z=z_0} = -A^{-1}(z) \left(\frac{d}{dz} A(z)\right) A^{-1}(z)\bigg|_{z=z_0}.$$

**Proof** When $A(z_0)$ is nonsingular, there is a nonempty open neighbourhood of $z_0$ over which $A^{-1}(z)$ is an entrywise differentiable function. Over this neighbourhood we have

$$A^{-1}(z) A(z) = I,$$

and so

$$\left(\frac{d}{dz}A^{-1}(z)\right)A(z) + A^{-1}(z)\left(\frac{d}{dz}A(z)\right) = 0.$$

Therefore,

$$\frac{d}{dz}A^{-1}(z) = -A^{-1}(z)\left(\frac{d}{dz}A(z)\right)A^{-1}(z)$$

at all points $z$ where $A(z)$ is nonsingular. ∎

A real matrix $A$ is positive definite if it is symmetric and every eigenvalue of $A$ is positive. We note that real positive definite matrices are nonsingular.

**Lemma A.9.** *Let $A$ be a real positive definite matrix and let $v$ be a nonzero real vector. Then,*

$$\left(v^T A v\right)\left(v^T A^{-1} v\right) \geq \left(v^T v\right)^2 = \|v\|^4,$$

*with equality if and only if $v$ is an eigenvector of $A$.*

**Proof** We make use of some well-known facts from linear algebra.

First, the Cauchy-Schwarz inequality (as it applies to real spaces of column vectors) is the following proposition: Let $v$ and $w$ be nonzero real column vectors, then,

$$v^T w \leq \|v\|\,\|w\| = (v^T v)^{1/2}(w^T w)^{1/2},$$

with equality if and only if $v = \beta w$ for some nonzero real number $\beta$.

Second, we make use of the following propositions, taken from [14, Chapter 7]: Let $A$ be a real positive definite matrix, then

1. there is a unique real positive definite matrix, labelled $A^{1/2}$, and referred to as the square root of $A$, such that

$$(A^{1/2})^2 = A;$$

2. the matrix $A^{-1}$ is itself real and positive definite; and

3. the square root of $A^{-1}$ is the inverse of the square root of $A$,

$$(A^{-1})^{1/2} = (A^{1/2})^{-1},$$

and we label this matrix $A^{-1/2}$.

(We have modified the results in [14] slightly, as we are only interested in the real case).

Now, let $A$ be a real positive definite matrix and let $v$ be a nonzero real vector. Then,

$$v^T v = v^T A^{1/2} A^{-1/2} v = (A^{1/2} v)^T (A^{-1/2} v).$$

So, via the Cauchy-Schwarz inequality,

$$v^T v \leq \left\| A^{1/2} v \right\| \left\| A^{-1/2} v \right\| = (v^T A v)^{1/2} (v^T A^{-1} v)^{1/2}.$$

Squaring every term in this expression obtains the expression in the above statement. Further, we note that equality holds if and only if

$$A^{1/2}v = \beta A^{-1/2}v,$$

for some real number $\beta$. When this occurs, multiplying both sides of this equality by $A^{1/2}$ obtains $Av = \beta v$. ∎

Let $B \in \mathcal{B}(n, \epsilon)$ and let $\Pi = \Pi_B$. We note that $\Pi_B = I$ if and only if $B$ is symmetric. If $B$ is symmetric, then

$$\alpha(B) = \min_{i \ni \pi_i = 1} \{\alpha_B(i)\} = \min\{\alpha_B(i)\}$$

and we have $\alpha_B(i) \geq \alpha(B)$ for all $i$. As well, if $B$ is symmetric, the expression of $B$ found in Lemma A.6 is

$$B \cong \begin{bmatrix} a & v^T \\ v & A \end{bmatrix},$$

where $A$ is symmetric and $\alpha(B) = a + v^T(I - A)^{-1}v$.

**Lemma A.10.** *Let $B \in \mathcal{B}(n, \epsilon)$. If $B$ is not symmetric, then there is a symmetric substochastic matrix $\hat{B} \in \mathcal{B}(n, \epsilon)$ such that $\alpha(\hat{B}) < \alpha(B)$.*

**Proof** Suppose that $B \in \mathcal{B}(n, \epsilon)$ is not symmetric. Express

$$B \cong \begin{bmatrix} a & v^T \\ w & A \end{bmatrix} \text{ and } \Pi = \Pi_B \cong \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix},$$

as in Lemma A.6. So, $Q \leq I$, $Qw = v$ and $QA$ is symmetric. The assumption that $B$ is not symmetric implies that $Q \neq I$. We note that since $B \in \mathcal{B}(n, \epsilon)$, we have

1. $B$ is irreducible,

2. $1 - \epsilon \leq a + v^T \mathbb{1} \leq 1$, and

3. $(1 - \epsilon)\mathbb{1} \leq A\mathbb{1} + w \leq \mathbb{1}$,

Let

$$\hat{B} = \begin{bmatrix} a & v^T \\ v & \hat{A} \end{bmatrix},$$

where

$$\hat{A} = QA + (1 - \epsilon)(I - Q).$$

We claim that $\hat{B}$ is a symmetric member of $\mathcal{B}(n, \epsilon)$ and $\alpha(\hat{B}) < \alpha(B)$. Since $0 \leq Q \leq I$, $\hat{B}$ is nonnegative. For every $i \neq j$, we have $\hat{b}_{ij} = \pi_i b_{ij}$; so, the fact that $B$ is irreducible implies that $\hat{B}$ is irreducible. As well, the fact that $QA$ is symmetric implies that $\hat{B}$ is symmetric. So, we next need to show that $\hat{B}$ is substochastic and $\gamma_{\hat{B}} \leq \epsilon\mathbb{1}$.

By assumption, $1 - \epsilon \leq a + v^T \mathbb{1} \leq 1$.

Next, $A\mathbb{1} + w \geq (1 - \epsilon)\mathbb{1}$ implies that

$$\begin{aligned} \hat{A}\mathbb{1} + v &= \left(QA + (1 - \epsilon)(I - Q)\right)\mathbb{1} + Qw \\ &= Q(A\mathbb{1} + w) + (1 - \epsilon)(I - Q)\mathbb{1} \\ &\geq Q\left((1 - \epsilon)\mathbb{1}\right) + (1 - \epsilon)(I - Q)\mathbb{1} \\ &= (1 - \epsilon)\mathbb{1}. \end{aligned}$$

As well, $A\mathbb{1} + w \leq \mathbb{1}$ and $0 \leq Q \leq I$ imply that

$$\begin{aligned}
\hat{A}\mathbb{1} + v &= Q(A\mathbb{1} + w) + (1-\epsilon)(I-Q)\mathbb{1} \\
&\leq Q\mathbb{1} + (1-\epsilon)(I-Q)\mathbb{1} \\
&= (1-\epsilon)\mathbb{1} + \epsilon Q\mathbb{1} \\
&\leq (1-\epsilon)\mathbb{1} + \epsilon\mathbb{1} \\
&= \mathbb{1}.
\end{aligned}$$

So, $(1-\epsilon)\mathbb{1} \leq \hat{B}\mathbb{1} \leq \mathbb{1}$.

Thus, $\hat{B}$ is a symmetric member of $\mathcal{B}(n,\epsilon)$. We now show that $\alpha(\hat{B}) < \alpha(B)$.

Since $\hat{B}$ is symmetric, $\Pi_{\hat{B}} = I$. So, $\alpha_{\hat{B}}(1) \geq \alpha(\hat{B})$. We note that $Qw = v$; thus, $w = Q^{-1}v$. We calculate

$$\begin{aligned}
\alpha(B) &= a + v^T (I-A)^{-1} w \\
&= a + v^T (I-A)^{-1} Q^{-1} v \\
&= a + v^T (Q - QA)^{-1} v
\end{aligned}$$

and

$$\begin{aligned}
\alpha_{\hat{B}}(1) &= a + v^T \left(I - \hat{A}\right)^{-1} v \\
&= a + v^T \left(I - (QA + (1-\epsilon)(I-Q))\right)^{-1} v \\
&= a + v^T \left(Q + \epsilon(I-Q) - QA\right)^{-1} v.
\end{aligned}$$

Permute the indices (if necessary) so that

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_l \end{bmatrix}, \; Q \cong \begin{bmatrix} Q_1 & & 0 \\ & \ddots & \\ 0 & & Q_l \end{bmatrix} \text{ and } v \cong \begin{bmatrix} v_1 \\ \vdots \\ v_l \end{bmatrix},$$

where each $A_k$ is irreducible. We expand our above formulae for $\alpha(B)$ and $\alpha_{\hat{B}}(1)$:

$$\alpha(B) = a + \sum_{k=1}^{l} v_k^T (Q_k - Q_k A_k)^{-1} v_k$$

$$\text{and } \alpha_{\hat{B}}(1) = a + \sum_{k=1}^{l} v_k^T \left(Q_k + \epsilon(I - Q_k) - Q_k A_k\right)^{-1} v_k.$$

If $Q_k = I$, the $k$th terms from the two sums are equal. If $Q_k \neq I$, we apply Lemma A.7 with $X = Q_k$, $Y = Q_k + \epsilon(I - Q_k)$ and $Z = Q_k A_k$ to see that, entrywise,

$$0 < \left(Q_k + \epsilon(I - Q_k) - Q_k A_k\right)^{-1} < (Q_k - Q_k A_k)^{-1}.$$

Since $B$ is irreducible, every $v_k$ has at least one positive term. Thus, if $Q_k \neq I$,

$$v_k^T \left(Q_k + \epsilon(I - Q_k) - Q_k A_k\right)^{-1} v_k < v_k^T (Q_k - Q_k A_k)^{-1} v_k.$$

Since $Q \neq I$, there is at least one $Q_k \neq I$ and so

$$\alpha(\hat{B}) \leq \alpha_{\hat{B}}(1) < \alpha(B).$$

∎

**Lemma A.11.** *Let $B \in \mathcal{B}(n, \epsilon)$ be symmetric. Suppose that there is an index $i$ such that $\alpha_B(i) = \alpha(B)$ and $\gamma_B(i) < \epsilon$. Then, there is a symmetric substochastic matrix $\hat{B} \in \mathcal{B}(n, \epsilon)$ such that*

1. *$\alpha(B) > \alpha(\hat{B})$,*

2. $\gamma_{\hat{B}}(i) = \epsilon$, *and*

3. *for all* $j \neq i$, $\gamma_{\hat{B}}(j) = \gamma_B(j)$.

**Proof** Let state $i$ be such that $\alpha_B(i) = \alpha(B)$ and $\gamma_B(i) < \epsilon$. Without loss of generality, we assume that $i = 1$. By assumption, $B$ is symmetric; via Lemma A.6, we express

$$B = \begin{bmatrix} a & v^T \\ v & A \end{bmatrix}$$

where $A$ is symmetric and $\alpha(B) = \alpha_B(1) = a + v^T(I - A)^{-1}v$. Now,

$$\gamma_B(1) = 1 - a - v^T\mathbb{1}.$$

So, $\gamma_B(1) < \epsilon$ implies that $a + v^T\mathbb{1} > 1 - \epsilon$.

First, suppose that $v^T\mathbb{1} \leq 1 - \epsilon$. Then, $a > 1 - \epsilon - v^T\mathbb{1} \geq 0$. Let

$$\hat{B} = \begin{bmatrix} 1 - \epsilon - v^T\mathbb{1} & v^T \\ v & A \end{bmatrix}.$$

We have

$$\begin{aligned} \alpha(\hat{B}) &\leq \alpha_{\hat{B}}(1) \\ &= 1 - \epsilon - v^T\mathbb{1} + v^T(I - A)^{-1}v \\ &< a + v^T(I - A)^{-1}v \\ &= \alpha(B). \end{aligned}$$

So, we next assume that $v^T\mathbb{1} > 1-\epsilon$. Let $R$ be the diagonal matrix with $r_i = v(i)$; thus, $R\mathbb{1} = v$. For real numbers $z$ with $0 \leq z < 1$, let $v(z) = (1-z)v$ and let $A(z) = A + zR$. As long as $z < 1$, the matrix $A(z)$ is properly substochastic, so $(1 - A(z))^{-1}$ is nonnegative. We will first show that the function

$$f(z) = v(z)^T(I - A(z))^{-1}v(z) = (1-z)^2 v^T(I - A(z))^{-1}v$$

is strictly decreasing in $z$ over the interval $z \in [0, 1)$. We note that

$$\frac{d}{dz}v(z) = -v \text{ and} \frac{d}{dz}A(z) = R.$$

Using Lemma A.8, we calculate

$$\begin{aligned}
\tfrac{df}{dz} &= (1-z)^2 v^T \left(\tfrac{d}{dz}(I - A(z))^{-1}\right) v \\
&\quad + \left(\tfrac{d}{dz}(1-z)^2\right) v^T(I - A(z))^{-1}v \\
&= (1-z)^2 v^T(I - A(z))^{-1} \left(-\tfrac{d}{dz}(I - A(z))\right)(I - A(z))^{-1}v \\
&\quad + \left(\tfrac{d}{dz}(1-z)^2\right) v^T(I - A(z))^{-1}v \\
&= (1-z)^2 v^T(I - A(z))^{-1}R(I - A(z))^{-1}v \\
&\quad -2(1-z)v^T(I - A(z))^{-1}v \\
&= v(z)^T(I - A(z))^{-1}R(I - A(z))^{-1}v(z) \\
&\quad -2v(z)^T(I - A(z))^{-1}v \\
&= v(z)^T(I - A(z))^{-1}R(I - A(z))^{-1}v(z) \\
&\quad -2v(z)^T(I - A(z))^{-1}R\mathbb{1} \\
&= v(z)^T(I - A(z))^{-1}R\left((I - A(z))^{-1}v(z)) - 2\mathbb{1}\right).
\end{aligned}$$

An application of Lemma A.7, together with the fact that $B$ is irreducible shows that the vector

$$v(z)^T(I - A(z))^{-1}R$$

is entrywise nonnegative with at least one positive entry (as long as $0 \leq z < 1$). We will show that the vector

$$(I - A(z))^{-1}v(z) - 2\mathbb{1}$$

has every entry negative. We note that

$$\begin{aligned} A(z)\mathbb{1} + v(z) &= A\mathbb{1} + zR\mathbb{1} + (1-z)v \\ &= A\mathbb{1} + zv + (1-z)v \\ &= A\mathbb{1} + v \\ &\leq \mathbb{1}. \end{aligned}$$

Thus, $v(z) \leq \mathbb{1} - A(z)\mathbb{1} = (I - A(z))\mathbb{1}$. This implies that

$$(I - A(z))^{-1}v(z) \leq (I - A(z))^{-1}(I - A(z))\mathbb{1} = \mathbb{1} < 2\mathbb{1},$$

and so

$$(I - A(z))^{-1}v(z) - 2\mathbb{1} < 0.$$

So, we have shown that $f(z) < f(0)$ as long as $0 < z < 1$. Let $z_0$ be such that

$$v(z_0)^T\mathbb{1} = (1 - z_0)v^T\mathbb{1} = 1 - \epsilon.$$

Since $1 - \epsilon < v^T \mathbb{1} \leq 1$, we have $0 < z_0 \leq \epsilon < 1$. Let

$$\hat{B} = \begin{bmatrix} 0 & v(z_0)^T \\ v(z_0) & A(z_0) \end{bmatrix}.$$

Since $v(z_0)$ is a positive scalar multiple of $v$ and $A(z_0)$ is equal to the sum of $A$ and a nonnegative diagonal matrix, $\hat{B}$ is an irreducible nonnegative matrix. The sum of the entries in the first row of $\hat{B}$ is $1 - \epsilon$ and the sum of the entries in any other row is equal to the sum of the entries in the corresponding row of $B$. Thus,

$$(1 - \epsilon)\mathbb{1} \leq \hat{B}\mathbb{1} \leq \mathbb{1}.$$

Finally, $A(z_0)$ is symmetric, since $A$ is symmetric. Thus, $\hat{B}$ is a symmetric member of $\mathcal{B}(n, \epsilon)$ and $Q_{\hat{B}} = I$. Then, we note that

$$\alpha(\hat{B}) \leq \alpha_{\hat{B}}(1) = f(z_0)$$

and

$$f(z_0) < a + f(0) = a + bv^T(I - A)^{-1}v = \alpha(B).$$

∎

Let $\mathcal{B} = \mathcal{B}(n, \epsilon)$. In calculating the value

$$\alpha(\mathcal{B}) = \inf_{B \in \mathcal{B}} \{\alpha(B)\},$$

it is sufficient to find a lower bound for $\alpha(B)$ where $B$ is a symmetric member of $\mathcal{B}$ (Lemma A.10), and $\alpha(B) = \alpha_B(i)$ where $\gamma_B(i) = \epsilon$ (Lemma A.11).

## A.3 A lower bound concerning stochastic complements of reversible substochastic matrices

We now calculate the value of

$$\alpha(\mathcal{B}) = \inf_{B \in \mathcal{B}} \{\alpha(B)\},$$

where $\mathcal{B} = \mathcal{B}(n, \epsilon)$. For $n = 1$, the problem is trivial. In this case $\mathcal{B} = \{[b] : 1 - \epsilon \leq b \leq 1\}$. For $B = [b] \in \mathcal{B}$, we have $\alpha(B) = b$; so, in this case,

$$\alpha(\mathcal{B}) = \inf_{B \in \mathcal{B}} \{\alpha(B)\} = 1 - \epsilon.$$

**Proposition A.12.** *Let $n$ be a positive integer greater than or equal to 2 and $\epsilon$ be a positive real number strictly less than 1; let $\mathcal{B} = \mathcal{B}(n, \epsilon)$. Then,*

$$\alpha(\mathcal{B}) = \frac{(1-\epsilon)^2}{1 + (n-2)\epsilon}.$$

*Moreover, a matrix $B \in \mathcal{B}$ has $\alpha(B) = \alpha(\mathcal{B})$ if and only if*

$$B \cong \begin{bmatrix} 0 & \frac{1-\epsilon}{n-1}\mathbb{1}^T \\ \frac{1-\epsilon}{n-1}\mathbb{1} & A \end{bmatrix},$$

*where $A$ is an $(n-1) \times (n-1)$ symmetric nonnegative matrix such that*

$$A\mathbb{1} = (1-\epsilon)\mathbb{1} - \frac{1-\epsilon}{n-1}\mathbb{1} = \frac{(1-\epsilon)(n-2)}{n-1}\mathbb{1}.$$

**Proof** By Lemmas A.6, A.10 and A.11, we simply have to calculate a lower bound for $a + v^T(I - A)^{-1}v$ where

1. the matrix $A$ is symmetric, nonnegative and has order $n-1$,

2. the vector $v$ is nonnegative, has order $n-1$ and satisfies $v^T\mathbb{1} \leq 1-\epsilon$,

3. the matrix

$$B = \begin{bmatrix} a & v^T \\ v & A \end{bmatrix},$$

   is substochastic and irreducible,

4. $a + v^T\mathbb{1} = 1-\epsilon$, and

5. $A\mathbb{1} + v \geq (1-\epsilon)\mathbb{1}$.

Let $A$, $v$ and $a$ satisfy the above and let $m = n-1 \geq 1$ be the order of $A$ and $v$. Let

$$r = A\mathbb{1} + v - (1-\epsilon)\mathbb{1};$$

we note that $r \geq 0$. Let $R$ be the diagonal matrix of order $m$ with $i$th diagonal entry equal to $r_i$. As in the proof of Lemma A.10, express

$$A \cong \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_l \end{bmatrix}, \; R \cong \begin{bmatrix} R_1 & & 0 \\ & \ddots & \\ 0 & & R_l \end{bmatrix} \text{ and } v \cong \begin{bmatrix} v_1 \\ \vdots \\ v_l \end{bmatrix},$$

where each $A_k$ is irreducible. As $B$ is irreducible, each $v_k$ has at least one positive entry. An application of Lemma A.7, with $Y = I + R_k$, $X = I$ and $Z = A_k$ shows that if $R_k \neq 0$, then the matrix $(I + R_k - A_k)^{-1}$ exists and, entrywise,

$$0 < (I + R_k - A_k)^{-1} < (I - A_k)^{-1}.$$

Thus, the matrix $(I + R - A)^{-1}$ is entrywise nonnegative. Let

$$\alpha' = a + v^T (I + R - A)^{-1} v.$$

Then,

$$\begin{aligned} \alpha' &= a + v^T (I + R - A)^{-1} v \\ &= a + \sum_{i=1}^{k} v_i^T (I + R_i - A_i)^{-1} v_i \\ &\leq a + \sum_{i=1}^{k} v_i^T (I - A_i)^{-1} v_i \\ &= \alpha(B), \end{aligned}$$

with equality if and only if $R = 0$. We note that $R = 0$ if and only if $B\mathbb{1} = (1 - \epsilon)\mathbb{1}$.

Now, let $A' = A - R$, so that

$$\alpha' = a + v^T (I - A')^{-1} v.$$

Although the matrix $A'$ may have negative entries, the matrix $(I - A')^{-1} = (I + R - A)^{-1}$ is entrywise nonnegative (as noted above). Since $R\mathbb{1} = A\mathbb{1} + v - (1 - \epsilon)\mathbb{1}$, we have $A'\mathbb{1} + v = (1 - \epsilon)\mathbb{1}$, implying that

$$v = (1 - \epsilon)\mathbb{1} - A'\mathbb{1} = (I - A')\mathbb{1} - \epsilon\mathbb{1}.$$

Thus,

$$\begin{aligned} v^T(I-A')^{-1}v &= \left(\mathbb{1}^T(I-A') - \epsilon\mathbb{1}^T\right)(I-A')^{-1}\left((I-A')\mathbb{1} - \epsilon\mathbb{1}\right) \\ &= \mathbb{1}^T(I-A')\mathbb{1} - 2\epsilon\mathbb{1}^T\mathbb{1} + \epsilon^2\mathbb{1}^T(I-A')^{-1}\mathbb{1}. \end{aligned}$$

As well,

$$\begin{aligned} a &= 1 - \epsilon - v^T\mathbb{1} \\ &= 1 - \epsilon - \left(\mathbb{1}^T(I-A') - \epsilon\mathbb{1}^T\right)\mathbb{1} \\ &= 1 - \epsilon - \mathbb{1}^T(I-A')\mathbb{1} + \epsilon\mathbb{1}^T\mathbb{1}. \end{aligned}$$

So,

$$\begin{aligned} \alpha' &= a + v^T(I-A')^{-1}v \\ &= 1 - \epsilon - \mathbb{1}^T(I-A')\mathbb{1} + \epsilon\mathbb{1}^T\mathbb{1} \\ &\quad + \mathbb{1}^T(I-A')\mathbb{1} - 2\epsilon\mathbb{1}^T\mathbb{1} + \epsilon^2\mathbb{1}^T(I-A')^{-1}\mathbb{1} \\ &= 1 - \epsilon - \epsilon\mathbb{1}^T\mathbb{1} + \epsilon^2\mathbb{1}^T(I-A')^{-1}\mathbb{1} \\ &= 1 - (m+1)\epsilon + \epsilon^2\mathbb{1}^T(I-A')^{-1}\mathbb{1}. \end{aligned}$$

(The vector $\mathbb{1}$ in the above expression has order $m$ and so $\mathbb{1}^T\mathbb{1} = m$). Thus, in order to calculate a lower bound for $\alpha'$ we simply need to calculate a lower bound for $\mathbb{1}^T(I-A')^{-1}\mathbb{1}$.

Now, $A$ and $A'$ are symmetric and $A - A' = R$, where $R$ is a positive semidefinite matrix ($R$ is a nonnegative diagonal matrix). The largest positive eigenvalue of $A'$ is less than or equal to the largest positive eigenvalue of $A$ (see [14, Corollary 7.7.4], for example). The matrix $A$ is properly substochastic, as it is a principal submatrix

of an irreducible substochastic matrix. The largest positive eigenvalue of $A$ is thus strictly less than 1. Altogether, $A'$ is a symmetric real matrix whose eigenvalues are strictly less than 1, further implying that $I - A'$ is a positive definite real matrix.

By Lemma A.9, we have

$$\left(\mathbb{1}^T(I - A')^{-1}\mathbb{1}\right)\left(\mathbb{1}^T(I - A')\mathbb{1}\right) \geq \left(\mathbb{1}^T\mathbb{1}\right)^2,$$

with equality if and only if $\mathbb{1}$ is an eigenvector of $A'$. Note that $A'\mathbb{1} + v = (1 - \epsilon)\mathbb{1}$ implies that $\mathbb{1}$ is an eigenvector of $A'$ if and only if $v$ is a scalar multiple of $\mathbb{1}$; so,

$$\mathbb{1}^T(I - A')^{-1}\mathbb{1} \geq \frac{(\mathbb{1}^T\mathbb{1})^2}{\mathbb{1}^T(I - A')\mathbb{1}} = \frac{m^2}{\mathbb{1}^T(I - A')\mathbb{1}},$$

with equality if and only if $v$ is a scalar multiple of $\mathbb{1}$. As well, $\mathbb{1}^T\mathbb{1} = m$ and

$$\begin{aligned}
\mathbb{1}^T(I - A')\mathbb{1} &= \mathbb{1}^T\mathbb{1} - \mathbb{1}^T A'\mathbb{1} &&= \mathbb{1}^T\mathbb{1} - \mathbb{1}^T\left((1 - \epsilon)\mathbb{1} - v\right) \\
&= \epsilon\mathbb{1}^T\mathbb{1} + v^T\mathbb{1} &&= m\epsilon + v^T\mathbb{1} \\
&\leq m\epsilon + (1 - \epsilon) &&= 1 + (m - 1)\epsilon.
\end{aligned}$$

(Recall that $v^T\mathbb{1} \leq 1 - \epsilon$.) Thus,

$$\mathbb{1}^T(I - A')^{-1}\mathbb{1} \geq \frac{m^2}{1 + (m - 1)\epsilon},$$

with equality if and only if $v$ is a scalar multiple of $\mathbb{1}$ and $v^T\mathbb{1} = 1 - \epsilon$. These two conditions uniquely identify $v$: when they both hold we have

$$v = \frac{1 - \epsilon}{m}\mathbb{1}.$$

So, in total, we have

$$\begin{aligned}
\alpha(B) &\geq \alpha' \\
&= 1-(m+1)\epsilon+\epsilon^2\mathbb{1}^T(I-A')^{-1}\mathbb{1} \\
&\geq 1-(m+1)\epsilon+\epsilon^2\tfrac{m^2}{1+(m-1)\epsilon} \\
&= \tfrac{(1-(m+1)\epsilon)(1+(m-1)\epsilon)+m^2\epsilon^2}{1+(m-1)\epsilon} \\
&= \tfrac{(1-\epsilon)^2}{1+(m-1)\epsilon},
\end{aligned}$$

with equality if and only if the matrix

$$B = \begin{bmatrix} a & v^T \\ v & A \end{bmatrix}$$

satisfies

1. $a + v^T\mathbb{1} = 1-\epsilon$,

2. $v = \frac{1-\epsilon}{m}\mathbb{1}$, and

3. $A\mathbb{1} + v = (1-\epsilon)\mathbb{1}$.

These three conditions together imply that $a = 0$ and $A\mathbb{1} = \frac{(1-\epsilon)(m-1)}{m}\mathbb{1}$. Substituting $m = n-1$ obtains the formulae in the statement of the proposition. ∎

Let $n \geq 1$, $\epsilon < 1$ and let $\mathcal{B} = \mathcal{B}(n,\epsilon)$. We note that the above formula for $\alpha(\mathcal{B})$ agrees with the case $n = 1$. As noted, when $n = 1$,

$$\alpha(\mathcal{B}) = 1-\epsilon = \frac{(1-\epsilon)^2}{1-\epsilon} = \frac{(1-\epsilon)^2}{1+(n-2)\epsilon}.$$

For $n = 1$ or 2, the matrices $B \in \mathcal{B}$ that have

$$\alpha(B) = \alpha(\mathcal{B}) = \frac{(1-\epsilon)^2}{1+(n-2)\epsilon} = (1-\epsilon)^2$$

are unique; they are

$$B = [1-\epsilon] \text{ or } B = \begin{bmatrix} 0 & 1-\epsilon \\ 1-\epsilon & 0 \end{bmatrix},$$

respectively.

However, this minimum for $\alpha(B)$ is not uniquely attained for $n \geq 3$. For example, the matrices

$$B_1 = \begin{bmatrix} 0 & \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & 0 \\ \frac{1-\epsilon}{2} & 0 & \frac{1-\epsilon}{2} \end{bmatrix}$$

$$\text{and } B_2 = \begin{bmatrix} 0 & \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & 0 & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & 0 \end{bmatrix}$$

satisfy

$$\alpha(B_1) = \alpha(B_2) = \alpha(\mathcal{B}) = \frac{(1-\epsilon)^2}{1+\epsilon}.$$

However, we can uniquely characterise those matrices that attain this value at every state.

**Proposition A.13.** *Let $B$ be an $n \times n$ substochastic matrix such that $\gamma_B \leq \epsilon$ and such that there is a positive diagonal matrix $\Pi$ with $\Pi B$ symmetric. Let*

$$p = \max_{1 \leq j \leq n} \{\pi_j\}.$$

*If $\pi_i = p$, then*

$$\alpha_B(i) \geq \frac{(1-\epsilon)^2}{1+(n-2)\epsilon}.$$

*Further,*

$$\alpha_B(j) = \frac{(1-\epsilon)^2}{1+(n-2)\epsilon}$$

*for all $j$ if and only if $n = 1$ and $B = [1-\epsilon]$, or $n \geq 2$ and*

$$B = \frac{1-\epsilon}{n-1}(J-I) = \begin{bmatrix} 0 & \frac{1-\epsilon}{n-1} & \cdots & \frac{1-\epsilon}{n-1} \\ \frac{1-\epsilon}{n-1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1-\epsilon}{n-1} \\ \frac{1-\epsilon}{n-1} & \cdots & \frac{1-\epsilon}{n-1} & 0 \end{bmatrix}.$$

**Proof** Let $\Pi$ be a positive diagonal matrix such that $\Pi B$ is symmetric and let $i$ be such that $\pi_i$ is maximal among the diagonal entries of $\Pi$.

First suppose that $[b_{ii}]$ is an irreducible block of $B$. This implies that the off-diagonal entries in the $i$th row and column of $B$ are 0. Since the sum of the entries in each row of $B$ is greater than or equal to $1-\epsilon$, we have $b_{ii} \geq 1-\epsilon$. So,

$$\alpha_B(i) = b_{ii} \geq 1-\epsilon = \frac{(1-\epsilon)^2}{1-\epsilon} \geq \frac{(1-\epsilon)^2}{1+(n-2)\epsilon},$$

with equality if and only $n = 1$ and $b_{ii} = 1 - \epsilon$.

Now, suppose that the $i$th row and column intersect an irreducible block of $B$ of order $n' \geq 2$. Let

$$\Pi' = \frac{1}{\pi_i}\Pi$$

and express

$$B \cong \begin{bmatrix} b_{ii} & v^T & 0 \\ w & B_1 & 0 \\ 0 & 0 & B_2 \end{bmatrix} \text{ and } \Pi' \cong \begin{bmatrix} 1 & 0 & 0 \\ 0 & \Pi_1 & 0 \\ 0 & 0 & \Pi_2 \end{bmatrix},$$

via the same similarity, where

$$\begin{bmatrix} b_{ii} & v^T \\ w & B_1 \end{bmatrix}$$

is irreducible. By Proposition A.12,

$$b_{ii} + v^T(I - B_1)w \geq \frac{(1-\epsilon)^2}{1 + (n'-2)\epsilon}.$$

We note that

$$\alpha_B(i) = b_{ii} + \begin{bmatrix} v^T & 0 \end{bmatrix} \begin{bmatrix} I - B_1 & 0 \\ 0 & I - B_2 \end{bmatrix}^{-1} \begin{bmatrix} w \\ 0 \end{bmatrix} = b_{ii} + v^T(I - B_1)^{-1}w.$$

Thus, since $n' \leq n$,

$$\alpha_B(i) \geq \frac{(1-\epsilon)^2}{1 + (n'-2)\epsilon} \geq \frac{(1-\epsilon)^2}{1 + (n-2)\epsilon}.$$

We note that we have equality if and only if $B$ is symmetric, $n' = n$, $\gamma_B = \epsilon\mathbb{1}$ and

$$v = w = \frac{1-\epsilon}{n-1}\mathbb{1}.$$

These conditions imply that $\Pi$ is a scalar multiple of the identity.

Suppose that we have

$$\alpha_B(j) = \frac{(1-\epsilon)^2}{1+(n-2)\epsilon}$$

for every $j$. Let $i$ be such that $\pi_i$ is maximal (among the diagonal of $\Pi$). Since

$$\alpha_B(i) = \frac{(1-\epsilon)^2}{1+(n-2)\epsilon},$$

our above reasoning implies that $\Pi$ is a scalar multiple of the identity and for all $j \neq i$, we have

$$b_{ij} = b_{ji} = \frac{1-\epsilon}{n-1}.$$

Thus, every index $j$ has $\pi_j$ maximal and so every off-diagonal entry of $B$ is equal to

$$\frac{1-\epsilon}{n-1}.$$

∎

The following proposition concerns the problem of finding a lower bound on the sum of the entries in a particular row of $B \setminus \mathcal{C}$, where $B \in \mathcal{B}(n, \epsilon)$.

**Proposition A.14.** *Let $B$ be a reversible substochastic matrix of order $n \geq 2$ such that $\gamma_B \leq \epsilon\mathbb{1}$. Let $\mathcal{E}$ be the state space of $B$ and let $\mathcal{C}$ be a nonempty subset of $\mathcal{E}$ containing $m \leq n-1$ states such that $\hat{B} = B \setminus \mathcal{C}$ exists. Let $\Pi$ be a positive diagonal matrix such that $\Pi B$ is symmetric and let*

$$p = \max_{j \in \mathcal{C}} \{\pi_j\}.$$

*Then, for all $i \in \mathcal{E} \setminus \mathcal{C}$ such that $\pi_i \geq p$,*

$$\sum_{j \in \mathcal{E} \setminus \mathcal{C}} \hat{b}_{ij} \geq \frac{(1-\epsilon)^2}{1+(m-1)\epsilon}.$$

**Proof** If $m = n-1$, then this is simply a restatement of Proposition A.12; so, we assume that $m \leq n-2$. Let $i \in \mathcal{E} \setminus \mathcal{C}$ be such that $\pi_i \geq \pi_j$ for all $j \in \mathcal{C}$. Express

$$B \cong \begin{bmatrix} b_{ii} & v_1^T & v_2^T \\ w_1 & A_1 & A_{12} \\ w_2 & A_{21} & A_2 \end{bmatrix},$$

where the first row and column correspond to state $i$, the second row and column of blocks corresponds to $\mathcal{E} \setminus (\mathcal{C} \cup \{i\})$ and the third row and column of blocks corresponds to $\mathcal{C}$. Without loss of generality, we assume that $\pi_i = 1$; so,

$$\Pi \cong \begin{bmatrix} 1 & 0 & 0 \\ 0 & Q_1 & 0 \\ 0 & 0 & Q_2 \end{bmatrix},$$

via the same permutation-similarity as $B$, and we have $Q_2 \leq I$.

The $i$th row of $B \setminus \mathcal{C}$ is equal to

$$\begin{bmatrix} b_{ii} + v_2^T(I - A_2)^{-1}w_2 & v_1^T + v_2^T(I - A_2)^{-1}A_{21} \end{bmatrix}.$$

So, we aim to show that

$$b_{ii} + v_2^T(I - A_2)^{-1}w_2 + v_1^T\mathbb{1} + v_2^T(I - A_2)^{-1}A_{21}\mathbb{1}$$

$$= b_{ii} + v_1^T\mathbb{1} + v_2^T(I - A_2)^{-1}(w_2 + A_{21}\mathbb{1}) \geq \frac{(1-\epsilon)^2}{1 + (m-1)\epsilon}.$$

Let $R$ be the nonnegative diagonal matrix of order $m$ that satisfies $R\mathbb{1} = A_{21}\mathbb{1}$. (The matrix $A_{21}$ has order $m \times (n - m - 1)$.) Consider the matrix

$$B' \cong \begin{bmatrix} b_{ii} + v_1^T\mathbb{1} & v_2^T \\ w_2 & A_2 + R \end{bmatrix}.$$

The matrix $B'$ is symmetrised by left-multiplication by the matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & Q_2 \end{bmatrix}.$$

Moreover, since $B\mathbb{1} \geq (1-\epsilon)\mathbb{1}$, we have

$$b_{ii} + v_1^T\mathbb{1} + v_2^T\mathbb{1} \geq 1 - \epsilon \text{ and } w_2 + A_{21}\mathbb{1} + A_2\mathbb{1} \geq (1-\epsilon)J.$$

By defining $R\mathbb{1} = A_{21}\mathbb{1}$, we have $B'\mathbb{1} \geq (1-\epsilon)\mathbb{1}$. So, by Proposition A.13 (and the fact that the order of $B'$ is $m + 1$), we have

$$b_{ii} + v_1^T \mathbb{1} + v_2^T (I - A_2 - R)^{-1} w_2 \geq \frac{(1-\epsilon)^2}{1 + (m-1)\epsilon}.$$

So, it is sufficient to show that

$$v_2^T (I - A_2)^{-1} (w_2 + A_{21}\mathbb{1}) \geq v_2^T (I - A_2 - R)^{-1} w_2.$$

Let

$$\gamma_2 = \mathbb{1} - A_2\mathbb{1} - A_{21}\mathbb{1} - w_2.$$

Since $B$ is substochastic, $\gamma \geq 0$. Further, since $R\mathbb{1} = A_{21}\mathbb{1}$, we have

$$w_2 = (I - A_2 - R)\mathbb{1} - \gamma_2 \text{ and } w_2 + A_{21}\mathbb{1} = (I - A_2)\mathbb{1} - \gamma_2.$$

Now, the fact that $R$ is nonnegative implies that

$$(I - A_2)^{-1} = \sum_{r=0}^{\infty} A_2^r \leq \sum_{r=0}^{\infty} (A_2 + R)^r = (I - A_2 - R)^{-1}.$$

So,

$$\begin{aligned} v_2^T (I - A_2 - R)^{-1} w_2 &= v_2^T (I - A_2 - R)^{-1} \left( (I - A_2 - R)\mathbb{1} - \gamma_2 \right) \\ &= v_2^T \mathbb{1} - v_2^T (I - A_2 - R)^{-1} \gamma_2 \\ &\leq v_2^T \mathbb{1} - v_2^T (I - A_2)^{-1} \gamma_2 \\ &= v_2^T (I - A_2)^{-1} \left( (I - A_2)\mathbb{1} - \gamma_2 \right) \\ &= v_2^T (I - A_2)^{-1} (w_2 + A_{21}\mathbb{1}). \end{aligned}$$

∎

# Appendix B

# A lower bound concerning stochastic complements of nonreversible Markov chains

## B.1 Preliminaries

Let $B$ be an irreducible substochastic matrix with state space $\mathcal{E}$ such that $\gamma_B \leq \epsilon \mathbb{1}$. Let $i \in \mathcal{E}$ and express

$$B \cong \begin{bmatrix} b_{ii} & v^T \\ w & A \end{bmatrix}$$

(as in Definition A.1). As in Appendix A, we aim to produce a lower bound on the term

$$\alpha_B(i) = b_{ii} + v^T (I - A)^{-1} w;$$

however, in this appendix, we do not assume that $B$ is reversible.

Let $X$ be a Markov chain with transition matrix $C$ and state space $\mathcal{S}$; let $\mathcal{E} \subseteq \mathcal{S}$ and $B = C(\mathcal{E})$. Recall that for $i \in \mathcal{S}$,

$$T_i = \inf\{t \geq 1 : x_t = i\}$$

is the first passage time into $i$ and that

$$E_{\mathcal{E}} = \inf\{t \geq 1 : x_t \notin \mathcal{E}\}.$$

Let $X$ be a Markov chain with transition matrix $A$ and let $B = A(\mathcal{E})$ be a principal submatrix corresponding to some proper subcollection of the state space; Proposition A.2 states that

$$\alpha_B(i) = \mathbb{P}_i\left[T_i < E_{\mathcal{E}}\right]$$

(when $\alpha_B(i)$ is defined). That is, $\alpha_B(i)$ is the probability that the Markov chain will transition from $i$ to $i$ (in 1 or more steps) without first exiting $\mathcal{E}$.

**Definition B.1.** Let $B$ be an irreducible substochastic matrix with state space $\mathcal{E}$. If $\mathcal{E}$ contains exactly two states, $i$ and $j$, we define

$$\alpha_B(i,j) = \frac{b_{ij}}{1 - b_{ii}} \text{ and } \alpha_B(j,i) = \frac{b_{ji}}{1 - b_{jj}}.$$

Suppose that $\mathcal{E}$ contains three or more states and let $i, j \in \mathcal{E}$ be distinct. Express

$$B \cong \begin{bmatrix} b_{ii} & b_{ij} & v_i^T \\ b_{ji} & b_{jj} & v_j^T \\ w_i & w_j & \tilde{B} \end{bmatrix},$$

where the first two positions correspond to $i$ and $j$ respectively and the remainder correspond to $\mathcal{E} \setminus \{i, j\}$. Then, we define

$$\alpha_B(i,j) = \frac{b_{ij} + v_i^T (I - \tilde{B})^{-1} w_j}{1 - b_{ii} - v_i^T (I - \tilde{B})^{-1} w_i}.$$

An alternate way to express the above definition is the following. Let $B$ be irreducible and substochastic with state space $\mathcal{E}$ and let $i, j \in \mathcal{E}$ be distinct; let $\hat{B} = B \setminus \{k : k \neq i \text{ or } j\}$. (If $\mathcal{E} = \{i, j\}$, then $\hat{B} = B \setminus \emptyset = B$.) Then,

$$\alpha_B(i,j) = \frac{\hat{b}_{ij}}{1 - \hat{b}_{ii}}.$$

**Proposition B.2.** *Let $X$ be an irreducible Markov chain with transition matrix $A$ and state space $\mathcal{S}$. Let $\mathcal{E} \subseteq \mathcal{S}$ contain two or more states, let $B = A(\mathcal{E})$ and let $i, j \in \mathcal{E}$ be distinct. Then,*

$$\alpha_B(i,j) = \mathbb{P}_i \left[ T_j < E_{\mathcal{E}} \right].$$

**Remark.** That is, for distinct $i, j \in \mathcal{E}$, $\alpha_B(i,j)$ is the probability that after visiting state $i$ the Markov chain will visit state $j$ at least once before exiting $\mathcal{E}$.

**Proof** Let $i$ and $j$ be distinct members of $\mathcal{E}$ and let $\mathcal{E}' = \mathcal{E} \setminus \{i, j\}$. Let $\hat{A} = A \setminus \mathcal{E}'$ and $\hat{B} = B \setminus \mathcal{E}'$. We note that $\hat{B} \cong \hat{A}(\{i, j\})$. In the same manner as in Proposition A.2, an application of Proposition 4.3 shows that

$$\hat{a}_{ii} = \hat{b}_{ii} = \mathbb{P}_i \left[ T_i < T_j \text{ and } T_i < E_{\mathcal{E}} \right]$$

is the probability of transitioning from $i$ to $i$ without first exiting $\mathcal{E}$ or visiting $j$. As well,

$$\hat{a}_{ij} = \hat{b}_{ij} = \mathbb{P}_i \left[ T_j < T_i \text{ and } T_j < E_{\mathcal{E}} \right]$$

is the probability of transitioning from $i$ to $j$ (in one or more steps) without transitioning into $i$ or exiting $\mathcal{E}$.

Thus, the probability, given $x_0 = i$, of transitioning into $i$ exactly $k \geq 0$ times before transitioning into $j$ for the first time, all without exiting $\mathcal{E}$, is $\hat{b}_{ii}^k \hat{b}_{ij}$.

If the Markov chain transitions from $i$ to $j$ without first exiting $\mathcal{E}$, before visiting $j$ for the first time it has visited $i$ some number $k \geq 0$ times. So,

$$\mathbb{P}_i \left[ T_j < E_{\mathcal{E}} \right] = \sum_{k \geq 0} \hat{b}_{ii}^k \hat{b}_{ij} = \frac{\hat{b}_{ij}}{1 - \hat{b}_{ii}}.$$

∎

Let $X$ be an irreducible Markov chain with state space $\mathcal{S}$ and transition matrix $A$. Let $\mathcal{E} \subseteq \mathcal{S}$ and $B = A(\mathcal{E})$; for $i, j \in \mathcal{E}$, we will take the value $\alpha_B(i, j)$ to be equal

to the probability that given $x_0 = i$, the Markov chain transitions into state $j$ at least once before exiting $\mathcal{E}$ for the first time:

$$\alpha_B(i,j) = \mathbb{E}_i \left[ T_j < E_{\mathcal{E}} \right].$$

Thus, we have $\alpha_B(i,i) = \alpha_B(i)$, as it appears in Definition A.1, and for $j \neq i$, $\alpha_B(i,j)$ is as in Definition B.1. Throughout this appendix, if the matrix $B$ is clearly specified, we will often use the labels $\alpha(i,j)$ and $\alpha(i)$ rather than $\alpha_B(i,j)$ and $\alpha_B(i)$; as well, we will use $\alpha(i)$ rather than $\alpha(i,i)$.

Let $X$ be a Markov chain on the state space $\mathcal{S}$. Recall that for each $i \in \mathcal{S}$ and $T \geq 1$,

$$N_i(T) = |\{t : 1 \leq t \leq T \text{ and } x_t = i\}|$$

is the total number of times the Markov chain has transitioned into $i$ at time $T$.

**Proposition B.3.** *Let $X$ be an Markov chain with state space $\mathcal{S}$ and transition matrix $A$. Let $\mathcal{E} \subseteq \mathcal{S}$ be such that $B = A(\mathcal{E})$ is properly substochastic. Then, for $i, j \in \mathcal{E}$ (not necessarily distinct),*

$$\mathbb{E}_i \left[ N_j(E_{\mathcal{E}}) \right] = \frac{\alpha_B(i,j)}{1 - \alpha_B(j)}.$$

**Remark.** That is, for $i, j \in \mathcal{E}$, we claim that, given $x_0 = i$, the expected number of transitions into $j$ before exiting $\mathcal{E}$ for the first time is $\alpha(i,j)/(1-\alpha(j))$.

**Proof** First, suppose that $\alpha(j) = 0$. Thus, it is impossible for the Markov chain to transition from $j$ to $j$ without exiting $\mathcal{E}$. So, if we have $x_0 = i$, then the Markov chain will transition into $j$ either once or no times at all before exiting $\mathcal{E}$. (If there are $k \geq 2$ visits to $j$ before exiting $\mathcal{E}$, then there must necessarily occur a transition $j \rightsquigarrow j$ without exiting $\mathcal{E}$.) The probability of one visit to $j$ before exiting $\mathcal{E}$ is $\alpha(i, j)$. So,

$$\begin{aligned}
\mathbb{E}_i \left[ N_j(E_{\mathcal{E}}) \right] &= \sum_{k \geq 1} k \mathbb{P}_i \left[ N_j(E_{\mathcal{E}}) = k \right] \\
&= \mathbb{P}_i \left[ N_j(E_{\mathcal{E}}) = 1 \right] \\
&= \alpha(i, j) \\
&= \tfrac{\alpha(i,j)}{1 - \alpha(j)}
\end{aligned}$$

(since, by assumption, $\alpha(j) = 0$).

Now, suppose that $\alpha(j) > 0$. Since $B = A(\mathcal{E})$ is properly substochastic, we must have $\alpha(j) < 1$ (it must be possible for the Markov chain to exit $\mathcal{E}$). As well, since

$$\alpha(j) = \mathbb{P}_j \left[ T_j < E_{\mathcal{E}} \right],$$

we have

$$1 - \alpha(j) = \mathbb{P}_j \left[ E_{\mathcal{E}} \leq T_j \right].$$

The fact that $j \in \mathcal{E}$ implies that $E_{\mathcal{E}} = T_j$ only if $E_{\mathcal{E}} = T_j = \infty$. Since $B = A(\mathcal{E})$ is properly substochastic, as we noted above, it must be possible for the Markov chain to exit $\mathcal{E}$; so, the probability that $E_{\mathcal{E}} = \infty$ is 0. Thus,

$$\begin{aligned} 1-\alpha(j) &= \mathbb{P}_j\left[E_{\mathcal{E}} \leq T_j\right] \\ &= \mathbb{P}_j\left[E_{\mathcal{E}} = T_j = \infty\right] + \mathbb{P}_j\left[E_{\mathcal{E}} < T_j\right] \\ &= \mathbb{P}_j\left[E_{\mathcal{E}} < T_j\right]. \end{aligned}$$

Suppose that $x_0 = i$ and that the Markov chain has transitioned into state $j$ exactly $k \geq 1$ times before exiting $\mathcal{E}$ for the first time. Then, the Markov chain

1. first transitioned from $i$ to $j$ without exiting $\mathcal{E}$, an event that has a probability of $\alpha(i,j)$ of occurring;

2. then transitioned from $j$ to $j$ without exiting $\mathcal{E}$ exactly $k-1$ times, events that each have a probability of $\alpha(j)$; and

3. the Markov chain then, starting from some $x_t = j$, exits $\mathcal{E}$ without visiting $j$ again – as we saw above, the probability of this occurring is $1-\alpha(j)$.

So, the probability that, given $x_0 = i$, the Markov chain transitions into $j$ exactly $k \geq 1$ times before exiting $\mathcal{E}$ for the first time is

$$\mathbb{P}_i\left[N_j(E_{\mathcal{E}}) = k\right] = \alpha(i,j)\alpha(j)^{k-1}(1-\alpha(j)).$$

In our next calculation, we take advantage of the well-known fact that for complex numbers $z$ with $0 < |z| < 1$,

$$\sum_{k\geq 1} k z^{k-1} = \frac{1}{(1-z)^2}.$$

Therefore, for $i, j \in \mathcal{E}$ (not necessarily distinct),

$$\begin{aligned}
\mathbb{E}_i\left[N_j(E_{\mathcal{E}})\right] &= \sum_{k\geq 1} k\mathbb{P}_i\left[N_j(E_{\mathcal{E}}) = k\right] \\
&= \sum_{k\geq 1} k\alpha(i,j)\alpha(j)^{k-1}(1-\alpha(j)) \\
&= \alpha(i,j)(1-\alpha(j))\sum_{k\geq 1} k\alpha(j)^{k-1} \\
&= \alpha(i,j)(1-\alpha(j))\tfrac{1}{(1-\alpha(j))^2} \\
&= \tfrac{\alpha(i,j)}{1-\alpha(j)}.
\end{aligned}$$

∎

**Lemma B.4.** *Let $B$ and $\hat{B}$ be substochastic matrices on the same state space $\mathcal{E}$ and suppose that $\hat{B} \leq B$. Then, for all $i, j \in \mathcal{E}$, not necessarily distinct, such that $\alpha_{\hat{B}}(i,j)$, $\alpha_B(i,j)$, $\alpha_{\hat{B}}(j)$ and $\alpha_B(j)$ are defined,*

$$\alpha_{\hat{B}}(i,j) \leq \alpha_B(i,j) \text{ and } \frac{\alpha_{\hat{B}}(i,j)}{1-\alpha_{\hat{B}}(j)} \leq \frac{\alpha_B(i,j)}{1-\alpha_B(j)}.$$

**Proof** We first show that $\alpha_{\hat{B}}(i) \leq \alpha_B(i)$. If the matrices in question have order 1, that is, if $\hat{B} = [\hat{b}_{ii}]$ and $B = [b_{ii}]$, then the statement is trivial: $\hat{B} \leq B$ implies that

$$\alpha_{\hat{B}}(i) = \hat{b}_{ii} \leq b_{ii} = \alpha_B(i).$$

Otherwise, express

$$B \cong \begin{bmatrix} b_{ii} & v^T \\ w & A \end{bmatrix} \text{ and } \hat{B} \cong \begin{bmatrix} \hat{b}_{ii} & \hat{v}^T \\ \hat{w} & \hat{A} \end{bmatrix}$$

via the same permutation-similarity. Thus, $\hat{b}_{ii} \leq b_{ii}$, $\hat{v} \leq v$, $\hat{w} \leq w$ and $\hat{A} \leq A$. We note that since $\hat{A} \leq A$,

$$(I - \hat{A})^{-1} = \sum_{k \geq 0} \hat{A}^k \leq \sum_{k \geq 0} A^k \leq (I - A)^{-1}.$$

So,

$$\begin{aligned} \alpha_{\hat{B}}(i) &= \hat{b}_{ii} + \hat{v}^T (I - \hat{A})^{-1} \hat{w} \\ &\leq b_{ii} + v^T (I - A)^{-1} w \\ &= \alpha_B(i). \end{aligned}$$

For $i \neq j$, we show that $\alpha_{\hat{B}}(i,j) \leq \alpha_B(i,j)$ in a very similar manner. Express

$$B \cong \begin{bmatrix} b_{ii} & b_{ij} & v_i^T \\ b_{ji} & b_{jj} & v_j^T \\ w_i & w_j & A \end{bmatrix} \text{ and } \hat{B} \cong \begin{bmatrix} \hat{b}_{ii} & \hat{b}_{ij} & \hat{v}_i^T \\ \hat{b}_{ji} & \hat{b}_{ij} & \hat{v}_i^T \\ \hat{w}_i & \hat{w}_j & \hat{A} \end{bmatrix}$$

via the same permutation-similarity. As before, $\hat{A} \leq A$ implies that $(I - \hat{A})^{-1} \leq (I - A)^{-1}$. So,

$$\hat{b}_{ij} + \hat{v}_i^T (I - \hat{A})^{-1} \hat{w}_j \leq b_{ij} + v_i^T (I - A)^{-1} w_j$$

and

$$\hat{b}_{ii} + \hat{v}_i^T (I - \hat{A})^{-1} \hat{w}_i \leq b_{ii} + v_i^T (I - A)^{-1} w_i.$$

This implies that

$$\alpha_{\hat{B}}(i,j) = \frac{\hat{b}_{ij} + \hat{v}_i^T (I - \hat{A})^{-1} \hat{w}_j}{1 - \hat{b}_{ii} - \hat{v}_i^T (I - \hat{A})^{-1} \hat{w}_i} \leq \frac{b_{ij} + v_i^T (I - A)^{-1} w_j}{1 - b_{ii} - v_i^T (I - A)^{-1} w_i} = \alpha_B(i,j).$$

The second inequality is a direct consequence of the first. The facts $\alpha_{\hat{B}}(i,j) \leq \alpha_B(i,j)$ and $\alpha_{\hat{B}}(j) \leq \alpha_B(j)$ imply that

$$\frac{\alpha_{\hat{B}}(i,j)}{1 - \alpha_{\hat{B}}(j)} \leq \frac{\alpha_B(i,j)}{1 - \alpha_B(j)}.$$

∎

## B.2 A lower bound concerning stochastic complements of substochastic matrices

Let $B$ be a substochastic matrix. Recall that

$$\gamma_B = (I - B)\mathbb{1},$$

where $\mathbb{1}$ is the column vector with every entry equal to 1, is a measure of how close $B$ is to being stochastic. Since

$$\gamma_B = \mathbb{1} - B\mathbb{1},$$

if $\gamma_B \leq \epsilon\mathbb{1}$ for some positive number $\epsilon \leq 1$, we have

$$B\mathbb{1} \geq (1 - \epsilon)\mathbb{1}.$$

**Proposition B.5.** *Let $\epsilon$ be a positive number strictly less than 1 and let $B$ be an irreducible substochastic matrix with state space $\mathcal{E}$ satisfying $\gamma_B = \epsilon\mathbb{1}$. Then, for all $i \in \mathcal{E}$,*

$$\sum_{j\in\mathcal{E}} \frac{\alpha(i,j)}{1-\alpha(j)} = \frac{1}{\epsilon} - 1 = \frac{1-\epsilon}{\epsilon}.$$

**Proof** Consider the following stochastic matrix,

$$C = \begin{bmatrix} 1 & 0 \\ \\ \epsilon\mathbb{1} & B \end{bmatrix}.$$

(Since $\gamma_B = \mathbb{1} - B\mathbb{1} = \epsilon\mathbb{1}$, we have $C\mathbb{1} = \mathbb{1}$.) We label the additional state not contained in $\mathcal{E}$ as state 0. Let $X$ be the Markov chain associated with $C$ on the state space $\mathcal{S} = \mathcal{E} \cup \{0\}$. We note that if $x_t \in \mathcal{E}$, then the probability that $x_{t+1} = 0$ is $\epsilon$ and the probability that $x_t \neq 0$ is $1-\epsilon$. Thus, given $x_0 \in \mathcal{E}$ and $t \geq 1$, the probability that $E_{\mathcal{E}} = t$ is $(1-\epsilon)^{t-1}\epsilon$.

So, for all $i \in \mathcal{E}$,

$$\mathbb{E}_i\left[E_{\mathcal{E}}\right] = \sum_{t\geq 1} t\mathbb{P}_i\left[E_{\mathcal{E}} = t\right] = \sum_{t\geq 1} t(1-\epsilon)^{t-1}\epsilon$$

$$= \epsilon \sum_{t\geq 1} t(1-\epsilon)^{t-1} = \epsilon\frac{1}{(1-(1-\epsilon))^2} = \frac{1}{\epsilon}.$$

The random variable $E_{\mathcal{E}}$ is the smallest $t \geq 1$ such that $x_t = 0$; so, the Markov chain transitions into states contained in $\mathcal{E}$ exactly $E_{\mathcal{E}} - 1$ times before exiting $\mathcal{E}$:

$$\sum_{j\in\mathcal{E}} N_j(E_{\mathcal{E}}) = E_{\mathcal{E}} - 1.$$

We apply Proposition B.3 to see that for all $i \in \mathcal{E}$,

$$\begin{aligned} \tfrac{1}{\epsilon} - 1 &= \mathbb{E}_i\left[E_{\mathcal{E}} - 1\right] \\ &= \sum_{j\in\mathcal{E}} \mathbb{E}_i\left[N_j(E_{\mathcal{E}})\right] \\ &= \sum_{j\in\mathcal{E}} \tfrac{\alpha(i,j)}{1-\alpha(j)}. \end{aligned}$$

∎

**Proposition B.6.** *Let $B$ be an irreducible substochastic matrix on states $\mathcal{E}$ such that $\gamma_B \leq \epsilon\mathbb{1}$. Then, for all $i \in \mathcal{E}$, we have*

$$\sum_{j\in\mathcal{E}} \frac{\alpha(i,j)}{1-\alpha(j)} \geq \frac{1-\epsilon}{\epsilon}.$$

**Proof** We note that since $B$ is irreducible and substochastic, each principal submatrix of $B$ that is not equal to $B$ itself is properly substochastic. Thus, $\alpha(i,j)$ is defined for any $i$ and $j$.

Let $R$ be the diagonal matrix whose $i$th diagonal entry is the sum of the entries in the $i$th row of $B$; so, $R\mathbb{1} = B\mathbb{1}$. Since $\gamma_B \leq \epsilon\mathbb{1}$, we have $(1-\epsilon)\mathbb{1} \leq R\mathbb{1} \leq \mathbb{1}$. Let

$$\hat{B} = (1-\epsilon)R^{-1}B.$$

Since each diagonal entry $r_i$ is greater than or equal to $1-\epsilon$,

$$\hat{b}_{ij} = \frac{1-\epsilon}{r_i} b_{ij} \leq b_{ij}.$$

Thus, $\hat{B} \leq B$ and

$$\hat{B}\mathbb{1} = (1-\epsilon)R^{-1}B\mathbb{1} = (1-\epsilon)\mathbb{1}.$$

By Lemma B.4 and Proposition B.5, for all $i \in \mathcal{E}$,

$$\sum_{j \in \mathcal{E}} \frac{\alpha_B(i,j)}{1-\alpha_B(j)} \geq \sum_{j \in \mathcal{E}} \frac{\alpha_{\hat{B}}(i,j)}{1-\alpha_{\hat{B}}(j)} = \frac{1-\epsilon}{\epsilon}.$$

∎

**Corollary B.7.** *Let $X$ be a nearly uncoupled Markov chain, with respect to $\epsilon > 0$, on the state space $\mathcal{S}$ and let $\mathcal{E} \subseteq \mathcal{S}$ be a minimal almost invariant aggregate. Then, for all $i \in \mathcal{E}$,*

$$\mathbb{E}_i[E_{\mathcal{E}}] \geq \frac{1}{\epsilon} \text{ and } \sum_{j \in \mathcal{E}} \mathbb{E}_i[N_j(E_{\mathcal{E}})] \geq \frac{1-\epsilon}{\epsilon}.$$

**Remark.** That is, if $\mathcal{E}$ is a minimal almost invariant aggregate, with respect to $\epsilon > 0$, of the Markov chain $X$, then, given $x_0 \in \mathcal{E}$, the expected value of the first exit time out of $\mathcal{E}$ is greater than or equal to $1/\epsilon$ and so the expected number of transitions into states contained in $\mathcal{E}$ before exiting $\mathcal{E}$ is greater than or equal to $1/\epsilon - 1 = (1-\epsilon)/\epsilon$.

A permutation matrix is a square $(0,1)$-matrix that has exactly one entry equal to 1 in each row and column. A cyclic permutation matrix is a permutation matrix $P$

whose associated digraph is a directed cycle. The cyclic permutation matrix of order 1 is simply $P = [1]$. For $n \geq 2$, the cyclic permutation matrices of order $n$ are those permutation matrices $P$ such that

$$P \cong \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{bmatrix}$$

(where the unspecified entries are zeroes). That is, the square matrix $P$ of order $n$ is a cyclic permutation matrix if there is an ordering of the integers

$$\{i_1, \ldots, i_n\} = \{1, \ldots, n\}$$

such that

$$p_{i_k i_l} = \begin{cases} 1 & \text{if } l \equiv k+1 \pmod{n} \\ 0 & \text{otherwise.} \end{cases}$$

**Proposition B.8.** *Let $B$ be a substochastic matrix on state space $\mathcal{E}$ such that $\gamma_B = \epsilon\mathbb{1}$. Then, for all $i \in \mathcal{E}$,*

$$\sum_{j \in \mathcal{E}} \alpha_B(i,j) \leq \sum_{m=1}^{n} (1-\epsilon)^m.$$

*Moreover, we have equality for one $i \in \mathcal{E}$ if and only if we have equality for every $i \in \mathcal{E}$, in which case $B = (1-\epsilon)P$, where $P$ is a cyclic permutation matrix.*

**Proof** If $n = 1$, then $\mathcal{E} = \{i\}$ and $B = [1 - \epsilon]$; so,

$$\sum_{j \in \mathcal{E}} \alpha_B(i, j) = \alpha_B(i, i) = 1 - \epsilon = \sum_{m=1}^{n} (1 - \epsilon)^m.$$

Thus, we assume that $n \geq 2$.

Without loss of generality, we assume that $\mathcal{E} = \{1, \ldots, n\}$. As in the proof of Proposition B.5, consider the stochastic matrix

$$C = \begin{bmatrix} 1 & 0 \\ \\ \epsilon \mathbb{1} & B \end{bmatrix}$$

and the Markov chain $X$ associated with $C$. The state space of $X$ is $\mathcal{S} = \mathcal{E} \cup \{0\}$, where we associated state 0 with the first column of $C$.

For $i \in \mathcal{E}$ and $m = 1, \ldots, n$, let

$$q(i, m) = \mathbb{P}_i \left[ |\{x_1, \ldots, x_{E_{\mathcal{E}}-1}\}| \geq m \right]$$

be the probability that, starting from $x_0 = i$, the Markov chain transitions into at least $m$ distinct members of $\mathcal{E}$ before exiting $\mathcal{E}$.

Since $B\mathbb{1} = (1 - \epsilon)\mathbb{1}$, for each $i \in \mathcal{E}$ and $T \geq 1$, the probability that

$$\{x_1, \ldots, x_T\} \subseteq \mathcal{E}$$

is equal to $(1 - \epsilon)^T$. Thus, for all $i \in \mathcal{E}$,

$$P_i \left[ E_{\mathcal{C}} > T \right] = (1 - \epsilon)^T.$$

We note that if

$$|\{x_1, \ldots, x_{E_{\mathcal{E}}-1}\}| \geq m,$$

then we must have $E_{\mathcal{E}} > m$. So, for all $i \in \mathcal{E}$ and $1 \leq m \leq n$,

$$q(i, m) \leq P_i \left[E_{\mathcal{C}} > m\right] = (1 - \epsilon)^m.$$

We claim that for each $i \in \mathcal{E}$,

$$\sum_{j \in \mathcal{E}} \alpha(i, j) = \sum_{m=1}^{n} q(i, m).$$

For each state $i \in \mathcal{E}$ and subset $\mathcal{C} \subseteq \mathcal{E}$, let

$$p(i, \mathcal{C}) = \mathbb{P}_i \left[\{x_1, \ldots, x_{E_{\mathcal{E}}-1}\} = \mathcal{C}\right].$$

That is, $p(i, \mathcal{C})$ is the probability that if $x_0 = i$, the members of $\mathcal{C}$ are exactly those states that the Markov chain transitions into at least once before exiting $\mathcal{E}$.

The number $\alpha(i, j)$ is the probability, given $x_0 = i$, of visiting $j$ before exiting $\mathcal{E}$. Thus,

$$\alpha(i, j) = \mathbb{P}_i \left[T_j < E_{\mathcal{E}}\right] = \mathbb{P}_i \left[j \in \{x_1, \ldots, x_{E_{\mathcal{E}}-1}\}\right],$$

implying that

$$\alpha(i, j) = \sum_{\substack{\mathcal{C} \subseteq \mathcal{E} \\ \ni j \in \mathcal{C}}} p(i, \mathcal{C}).$$

Thus,

$$\sum_{j\in\mathcal{E}}\alpha(i,j)=\sum_{j\in\mathcal{E}}\sum_{\substack{\mathcal{C}\subseteq\mathcal{E}\\ \ni j\in\mathcal{C}}}p(i,\mathcal{C})=\sum_{\mathcal{C}\subseteq\mathcal{E}}\sum_{j\in\mathcal{C}}p(i,\mathcal{C})=\sum_{\mathcal{C}\subseteq\mathcal{E}}|\mathcal{C}|p(i,\mathcal{C}).$$

As well, it is clear that

$$q(i,m)=\mathbb{P}_i\left[|\{x_1,\ldots,x_{E_{\mathcal{E}}-1}\}|\geq m\right]=\sum_{\substack{\mathcal{C}\subseteq\mathcal{S}\\ \ni|\mathcal{C}|\geq m}}p(i,\mathcal{C}).$$

So,

$$\sum_{m=1}^{n}q(i,m)=\sum_{m=1}^{n}\sum_{\substack{\mathcal{C}\subseteq\mathcal{S}\\ \ni|\mathcal{C}|\geq m}}p(i,\mathcal{C})=\sum_{\mathcal{C}\subseteq\mathcal{E}}\sum_{m=1}^{|\mathcal{C}|}p(i,\mathcal{C})=\sum_{\mathcal{C}\subseteq\mathcal{E}}|\mathcal{C}|p(i,\mathcal{C}).$$

Therefore, for each $i\in\mathcal{E}$,

$$\sum_{j\in\mathcal{E}}\alpha(i,j)=\sum_{\mathcal{C}\subseteq\mathcal{E}}|\mathcal{C}|p(i,\mathcal{C})=\sum_{m=1}^{n}q(i,m).$$

As we noted above, $q(i,m)\leq(1-\epsilon)^m$, so, for all $i\in\mathcal{E}$,

$$\sum_{j\in\mathcal{E}}\alpha(i,j)\leq\sum_{m=1}^{n}(1-\epsilon)^m.$$

Now, suppose that for some $i\in\mathcal{E}$,

$$\sum_{j\in\mathcal{E}}\alpha(i,j)=\sum_{m=1}^{n}(1-\epsilon)^m.$$

Via our above reasoning, we must have $q(i,m)=(1-\epsilon)^m$ for $m=1,\ldots,n$. The number $q(i,m)$ is the probability that, given $x_0=i$, the Markov chain transitions into at least $m$ distinct states in $\mathcal{E}$ before exiting $\mathcal{E}$; the number $(1-\epsilon)^m$ is the probability

that the Markov chain transitions into at least $m$ states in $\mathcal{E}$, not necessarily distinct, before exiting $\mathcal{E}$. Thus, in this case, whenever the Markov chain begins at $x_0 = i$ and remains in $\mathcal{E}$ for $m$ transitions, where $1 \leq m \leq n$, the states $x_1, \ldots, x_m$ are distinct members of $\mathcal{E}$.

Let $G$ be the digraph on $\mathcal{E}$ associated with $B$. Since, starting from $x_0 = i$, it is possible to remain in $\mathcal{E}$ for $n$ transitions (namely, the probability of this occurring is $(1-\epsilon)^n > 0$), there is a directed walk in $G$ of length $n$ with initial vertex $i$:

$$\omega = i \to i_1 \to \cdots \to i_n.$$

Our above observations imply that the states $i_1, \ldots, i_n$ are distinct, so

$$\{i_1, \ldots, i_n\} = \{1, \ldots, n\}.$$

First, we claim that $i_n = i$. If we suppose not, then $i_m = i$ for some $m \leq n-1$. This would imply the presence of the directed walk

$$i \to i_1 \to \cdots \to i_m \to i_1$$

in $G$. This directed walk has length $m+1 \leq n$ and contains the state $i_1$ twice. Its presence in $G$ implies that it is possible for the Markov chain to remain in $\mathcal{E}$ for $m+1 \leq n$ steps and yet visit $m$ or fewer distinct states. This contradicts our assumptions and so it must be that $i_n = i$. Thus, $\omega$ is a directed cycle of length $n$ in $G$.

Now, suppose that $G$ contains a directed cycle with length $m \leq n-1$:

$$\upsilon = j_1 \to \cdots \to j_m \to j_1$$

(in a directed cycle of length $m$, the $m$ vertices present are distinct).

We first note that $i$ cannot appear in $\upsilon$. If $i$ were present in this directed cycle, it would be possible for the Markov chain to begin at $i$ and then transition into $m+1 \leq n$ members of $\mathcal{E}$ but only transition into $m$ distinct members of $\mathcal{E}$ (simply by following the transitions in $\upsilon$).

Let $k$ be the smallest index such that $i_k$ (using the labelling of $\omega$) appears in $\upsilon$. Let $l$ be the index of $i_k$ in $\upsilon$; *i.e.* let $l$ be such that $i_k = j_l$. So, since $i_n = i$ and $i$ does not appear in $\upsilon$, we have

$$\{j_1, \ldots, j_m\} \subseteq \{i_k, \ldots, i_{n-1}\},$$

implying that $m \leq n-k$. So, the existence of $\upsilon$ implies the presence of the following directed walk in $G$:

$$i \to i_1 \to \cdots \to i_k = j_l \to \cdots \to j_l,$$

where the transition $j_l \rightsquigarrow j_l$ is achieved by following $\upsilon$. This directed walk has initial vertex $i$, visits the state $j_l$ twice and has length $k+m \leq n$. This is a contradiction and so $G$ does not contain any directed cycles of length less than $n$.

It must be that $\omega$ contains every directed arc in $G$. The subgraph $\omega$ is a directed

cycle on all the vertices of $G$; if $G$ contains even one more directed arc, it must contain a directed cycle of length strictly less than $n$.

So, we find that if there is a state $i \in \mathcal{E}$ such that

$$\sum_{j \in \mathcal{E}} \alpha(i,j) = \sum_{m=1}^{n} (1-\epsilon)^m,$$

then

$$B \cong \begin{bmatrix} 0 & b_{i_1 i_2} & & \\ & \ddots & \ddots & \\ & & \ddots & b_{i_{n-1} i_n} \\ b_{i_n i_1} & & & 0 \end{bmatrix},$$

where every unspecified entry is 0. Since $B\mathbb{1} = (1-\epsilon)\mathbb{1}$, all of the nonzero entries in $B$ must be equal to $1-\epsilon$; thus, $B = (1-\epsilon)P$, where $P$ is a cyclic permutation, when equality holds for at least one $i \in \mathcal{E}$ (in the inequality in the statement).

If we suppose that $B = (1-\epsilon)P$, where $P$ is a cyclic permutation, then there is an ordering of $\mathcal{C}$ into $i_1, \ldots, i_n$ such that $b_{i_k i_{k+1}} = 1-\epsilon$ for $k = 1, \ldots, n-1$, $b_{i_n i_1} = 1-\epsilon$ and every other entry is 0. In a sense, the Markov chain associated with $C$ (above) is deterministic, if $x_0 = i_k$ and $x_t \neq 0$, then it must be that $x_t = i_l$ where $l \equiv k + t \pmod{n}$. Thus,

$$\alpha(i_k, i_l) = (1-\epsilon)^m$$

where $m$ is the unique positive integer less than or equal to $n$ such that $m \equiv l - k \pmod{n}$. It is clear that for all $i \in \mathcal{E}$,

$$\sum_{j\in\mathcal{E}} \alpha(i,j) = \sum_{m=1}^{n} (1-\epsilon)^m.$$

∎

**Proposition B.9.** *Let $B$ be an irreducible substochastic matrix such that $\gamma_B \leq \epsilon\mathbb{1}$, where $0 < \epsilon < 1$. Let $\mathcal{E}$ be the associated state space and let $n$ be the order of $\mathcal{C}$ (and $B$). Then, there is $i \in \mathcal{E}$ such that*

$$\alpha_B(i) \geq (1-\epsilon)^n.$$

*Moreover,*

$$\max_{i\in\mathcal{E}}\{\alpha_B(i)\} = (1-\epsilon)^n$$

*if and only if $B = (1-\epsilon)P$, where $P$ is a cyclic permutation matrix.*

**Proof** We first show that the proposition holds under the assumption that $\gamma_B = \epsilon\mathbb{1}$. Then, in a manner similar to Proposition B.6, we will show that this implies the proposition for substochastic matrices $B$ with $\gamma_B \leq \epsilon\mathbb{1}$.

Fix a specific state $i \in \mathcal{E}$. Via Proposition B.5, we have

$$\sum_{j\in\mathcal{E}} \frac{\alpha(i,j)}{1-\alpha(j)} = \frac{1-\epsilon}{\epsilon}.$$

As well, by Proposition B.8,

$$\sum_{j \in \mathcal{E}} \alpha_B(i,j) \leq \sum_{m=1}^{n} (1-\epsilon)^m,$$

with equality if and only if $B = (1-\epsilon)P$, where $P$ is a cyclic permutation matrix.

We will take advantage of the well-known fact that for any complex number $z \neq 0$ and positive integer $n$,

$$1 - z^n = (1-z)(1 + z + \ldots + z^{n-1}) = (1-z) \sum_{m=0}^{n-1} z^m.$$

This implies that

$$\begin{aligned} \tfrac{1-\epsilon}{\epsilon} \left(1 - (1-\epsilon)^n\right) &= \tfrac{1-\epsilon}{\epsilon} (1 - (1-\epsilon)) \sum_{m=0}^{n-1} (1-\epsilon)^m \\ &= (1-\epsilon) \sum_{m=0}^{n-1} (1-\epsilon)^m \\ &= \sum_{m=1}^{n} (1-\epsilon)^m. \end{aligned}$$

So, we have

$$\begin{aligned} \sum_{j \in \mathcal{E}} \alpha(i,j) \tfrac{(1-(1-\epsilon)^n)}{1-\alpha(j)} &= (1 - (1-\epsilon)^n) \sum_{j \in \mathcal{E}} \tfrac{\alpha(i,j)}{1-\alpha(j)} \\ &= (1 - (1-\epsilon)^n) \tfrac{1-\epsilon}{\epsilon} \\ &= \sum_{m=1}^{n} (1-\epsilon)^m \\ &\geq \sum_{j \in \mathcal{E}} \alpha(i,j). \end{aligned}$$

It must be that for at least one $j \in \mathcal{E}$,

$$\frac{1 - (1-\epsilon)^n}{1 - \alpha(j)} \geq 1,$$

which, in turn, implies that for at least one $j \in \mathcal{E}$, $\alpha(j) \geq (1-\epsilon)^n$.

Now, suppose that

$$\max_{j \in \mathcal{E}} \{\alpha(j)\} = (1-\epsilon)^n.$$

This implies that

$$1 \geq \frac{1-(1-\epsilon)^n}{1-\alpha(j)}$$

for all $j \in \mathcal{E}$. This, in turn, implies that

$$\sum_{j \in \mathcal{E}} \alpha(i,j) \geq (1-(1-\epsilon)^n) \sum_{j \in \mathcal{E}} \frac{\alpha(i,j)}{1-\alpha(j)} = \sum_{m=1}^{n} (1-\epsilon)^m.$$

By Proposition B.8, we have

$$\sum_{j \in \mathcal{E}} \alpha(i,j) \leq \sum_{m=1}^{n} (1-\epsilon)^m,$$

with equality if and only if $B = (1-\epsilon)P$, where $P$ is a cyclic permutation matrix. The above two inequalities clearly, together, imply that

$$\sum_{j \in \mathcal{E}} \alpha(i,j) = \sum_{m=1}^{n} (1-\epsilon)^m$$

and thus $B = (1-\epsilon)P$, where $P$ is a cyclic permutation.

Now, we consider the case that $\gamma_B \leq \epsilon \mathbb{1}$. Let $R$ be the diagonal matrix that satisfies $R\mathbb{1} = B\mathbb{1}$ and let $\hat{B} = (1-\epsilon)R^{-1}B$. As in the proof of Proposition B.6, we have $\hat{B} \leq B$ and so, via Lemma B.4, for all $i, j \in \mathcal{E}$ (not necessarily distinct), $\alpha_{\hat{B}}(i,j) \leq \alpha_B(i,j)$. As we have shown above,

$$\max_{i\in\mathcal{E}}\{\alpha_{\hat{B}}(i)\} \geq (1-\epsilon)^n,$$

with equality if and only if

$$(1-\epsilon)R^{-1}B = \hat{B} = (1-\epsilon)P,$$

where $P$ is a cyclic permutation matrix. Since

$$\max_{i\in\mathcal{E}}\{\alpha_{B}(i)\} \geq \max_{i\in\mathcal{E}}\{\alpha_{\hat{B}}(i)\},$$

we have

$$\max_{i\in\mathcal{E}}\{\alpha_{B}(i)\} \geq (1-\epsilon)^n.$$

If we suppose that equality occurs, then we must also have equality for $\hat{B}$ and so it must be that $R^{-1}B$ is a cyclic permutation matrix. This implies that

$$B \cong \begin{bmatrix} 0 & b_1 & & \\ & \ddots & \ddots & \\ & & \ddots & b_{n-1} \\ b_n & & & 0 \end{bmatrix}.$$

Then, each $b_k$ is greater than or equal to $1-\epsilon$ (since $B\mathbb{1} \geq (1-\epsilon)\mathbb{1}$). It is clear that for all $i \in \mathcal{E}$ the probability of transitioning from $i$ to $i$ is

$$\alpha_B(i) = \prod_{m=1}^{n} b_m \geq (1-\epsilon)^n.$$

Thus,

$$\max_{i\in\mathcal{E}}\{\alpha_B(i)\} = (1-\epsilon)^n$$

implies that we have $b_m = 1-\epsilon$ for all $m$ and so implies that $B = \hat{B} = (1-\epsilon)P$, where $P$ is a cyclic permutation matrix.

If $B = (1-\epsilon)P$ where $P$ is a cyclic permutation, then $\alpha_B(i) = (1-\epsilon)^n$ for all $i$ and so

$$\max_{i\in\mathcal{E}}\{\alpha_B(i)\} = (1-\epsilon)^n.$$

∎

**Remark.** Let $B$ be an irreducible substochastic matrix with state space $\mathcal{E}$ containing $n$ states; suppose that $\gamma_B \leq \epsilon\mathbb{1}$. We note that Proposition B.9 does not merely imply that at least one member $i \in \mathcal{E}$ has $\alpha(i) \geq (1-\epsilon)^n$. Within the proof, we see that there is, in fact, a family of weighted averages of the terms

$$\frac{1-(1-\epsilon)^n}{1-\alpha(i)}$$

that are each greater than or equal to 1. Namely, for each pair $i, j \in \mathcal{E}$, let

$$\beta(i,j) = \frac{\alpha(i,j)}{\sum\limits_{k\in\mathcal{E}}\alpha(i,k)}.$$

Then, for all $i \in \mathcal{E}$,

$$\sum_{j\in\mathcal{E}}\beta(i,j) = 1 \text{ and } \sum_{j\in\mathcal{E}}\beta(i,j)\frac{1-(1-\epsilon)^n}{1-\alpha(j)} \geq 1.$$

Thus, we expect that, on average, randomly chosen members $i \in \mathcal{E}$ have $\alpha(i) \geq (1-\epsilon)^n$.

## B.3 A lower bound concerning scalar multiples of doubly stochastic matrices

We examine the problem of finding a lower bound on $\alpha_B$, where $B$ is a scalar multiple of a doubly stochastic matrix.

A doubly stochastic matrix is a nonnegative square matrix $C$ such that $C\mathbb{1} = C^T\mathbb{1} = \mathbb{1}$. That is, the sum of the entries in any row or column of a doubly stochastic matrix is 1.

Let $B$ be substochastic matrix that is a scalar multiple of a doubly stochastic matrix. Then, we have $B\mathbb{1} = B^T\mathbb{1} = z\mathbb{1}$ for some real number $z$ with $0 \leq z \leq 1$. We are interested in the case that $B$ is a principal submatrix of a stochastic matrix corresponding to an almost invariant aggregate; thus, when a substochastic matrix $B$ has constant row and column sums, we will express $B = (1-\epsilon)C$ where $C$ is doubly stochastic.

In Lemma B.10, we use the convention that $A^0 = I$, for any real matrix $A$.

**Lemma B.10.** *Let $B = (1-\epsilon)C$ where $0 < \epsilon < 1$ and $C$ is doubly stochastic. Suppose that $B$ has order $n \geq 2$ and let $A$ be a principal submatrix of $B$ of order $n-1$. Then, for $k = 0, \ldots, n-2$,*

$$\mathbb{1}^T A^k \mathbb{1} \geq (n-k-1)(1-\epsilon)^k.$$

**Proof** We proceed by induction on $k$. For $k = 0$, the statement is trivial. The vector of all ones in question, $\mathbb{1}$, has order $n-1$ and so

$$\mathbb{1}^T A^0 \mathbb{1} = \mathbb{1}^T \mathbb{1} = n-1 = (n-0-1)(1-\epsilon)^0.$$

Express

$$B \cong \begin{bmatrix} a & v^T \\ w & A \end{bmatrix}.$$

Since $B = (1-\epsilon)C$, where $C$ is doubly stochastic, we have

$$A\mathbb{1} + w = (1-\epsilon)J \text{ and } \mathbb{1}^T w = 1 - \epsilon - a \leq 1 - \epsilon.$$

We assume that $1 \leq k \leq n-2$ and that

$$\mathbb{1}^T A^{k-1} \mathbb{1} \geq (n-(k-1)-1)(1-\epsilon)^{k-1}.$$

This hypothesis, together with the fact that $A\mathbb{1} + w = (1-\epsilon)\mathbb{1}$, implies that

$$\begin{aligned} \mathbb{1}^T A^k \mathbb{1} + \mathbb{1}^T A^{k-1} w &= \mathbb{1}^T A^{k-1} (A\mathbb{1} + w) \\ &= \mathbb{1}^T A^{k-1} ((1-\epsilon)\mathbb{1}) \\ &= (1-\epsilon)\mathbb{1}^T A^{k-1} \mathbb{1} \\ &\geq (1-\epsilon)(n-(k-1)-1)(1-\epsilon)^{k-1} \\ &= (n-k)(1-\epsilon)^k. \end{aligned}$$

So,

$$\mathbb{1}^T A^k \mathbb{1} \geq (n-k)(1-\epsilon)^k - \mathbb{1}^T A^{k-1} w.$$

Since $\mathbb{1}^T A + v^T = (1-\epsilon)\mathbb{1}^T$, we have $\mathbb{1}^T A \leq (1-\epsilon)\mathbb{1}^T$; this implies that $\mathbb{1}^T A^{k-1} \leq (1-\epsilon)^{k-1}\mathbb{1}^T$. So, since $\mathbb{1}^T w \leq 1-\epsilon$ and the vector $\mathbb{1}$ in the above inequalities has order $n-1$,

$$\mathbb{1}^T A^{k-1} w \leq (1-\epsilon)^{k-1}\mathbb{1}^T w \leq (1-\epsilon)^k.$$

Therefore,

$$\begin{aligned} \mathbb{1}^T A^k \mathbb{1} &\geq (n-k)(1-\epsilon)^k - \mathbb{1}^T A^{k-1} w \\ &\geq (n-k)(1-\epsilon)^k - (1-\epsilon)^k \\ &= (n-k-1)(1-\epsilon)^k. \end{aligned}$$

∎

**Lemma B.11.** *Let $\epsilon$ be a positive real number strictly less than 1 and let $n$ be a positive integer greater than or equal to 2. Then,*

$$\sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k = \frac{(1-\epsilon)^n - (1-n\epsilon)}{\epsilon^2}.$$

**Proof** We will proceed by induction on $n$. For $n = 2$, we have

$$\sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k = (1-\epsilon)^0 = 1.$$

As well

$$\frac{(1-\epsilon)^2-(1-2\epsilon)}{\epsilon^2} = \frac{1-2\epsilon+\epsilon^2-1+2\epsilon}{\epsilon^2} = 1.$$

So, we assume that $n \geq 2$ and that

$$\sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k = \frac{(1-\epsilon)^n-(1-n\epsilon)}{\epsilon^2}.$$

We aim to show that

$$\sum_{k=0}^{n-1}(n-k)(1-\epsilon)^k = \frac{(1-\epsilon)^{n+1}-(1-(n+1)\epsilon)}{\epsilon^2}$$

(thus proving that when the statement holds for $n$, it also holds for $n' = n+1$). For a real number $z \neq 0, 1$ and a positive integer $r \geq 1$,

$$\sum_{k=0}^{r-1} z^k = 1 + z + \ldots + z^{r-1} = \frac{1-z^r}{1-z}.$$

So, for $z = 1-\epsilon$ and $r = n$, this implies that

$$\sum_{k=0}^{n-1}(1-\epsilon)^k = \frac{1-(1-\epsilon)^n}{1-(1-\epsilon)} = \frac{1-(1-\epsilon)^n}{\epsilon}.$$

We calculate

$$\begin{aligned}
\sum_{k=0}^{n-1}(n-k)(1-\epsilon)^k &= \sum_{k=0}^{n-1}(n-k-1)(1-\epsilon)^k + \sum_{k=0}^{n-1}(1-\epsilon)^k \\
&= \sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k + \sum_{k=0}^{n-1}(1-\epsilon)^k \\
&= \frac{(1-\epsilon)^n-(1-n\epsilon)}{\epsilon^2} + \frac{1-(1-\epsilon)^n}{\epsilon} \\
&= \frac{(1-\epsilon)^n-(1-n\epsilon)+\epsilon-(1-\epsilon)^n\epsilon}{\epsilon^2} \\
&= \frac{(1-\epsilon)^{n+1}-(1-(n+1)\epsilon)}{\epsilon^2}.
\end{aligned}$$

∎

**Proposition B.12.** *Let $B = (1-\epsilon)C$ where $C$ is doubly stochastic and $0 < \epsilon < 1$. Let $\mathcal{E}$ be the associated state space and let $n$ be the order of $B$ and $\mathcal{E}$. Then, for all $i \in \mathcal{E}$,*

$$\alpha_B(i) \geq (1-\epsilon)^n.$$

*Moreover, equality is attained for at least one $i \in \mathcal{E}$ if and only if $C$ is a cyclic permutation matrix, in which case equality is attained for every $i \in \mathcal{E}$.*

**Proof** We note that if $n = 1$, the statement is trivial. In this case we have $B = [1-\epsilon]$, $\mathcal{E} = \{i\}$ and $\alpha_B(i) = 1-\epsilon$. So, we assume that $n \geq 2$. Reordering the states $\mathcal{E}$ does not alter the fact that $B$ is a scalar multiple of a doubly stochastic matrix; so, we will simply show that $\alpha_B(1) \geq (1-\epsilon)^n$ with equality if and only $C$ is a cyclic permutation matrix.

Express

$$B = \begin{bmatrix} a & v^T \\ w & A \end{bmatrix}.$$

The fact that $B\mathbb{1} = B^T\mathbb{1} = (1-\epsilon)\mathbb{1}$ implies that

$$a = 1-\epsilon-v^T\mathbb{1},\ w = (I-A)\mathbb{1} - \epsilon\mathbb{1} \text{ and } v = (I-A)^T\mathbb{1} - \epsilon\mathbb{1}.$$

So, we calculate

$$\begin{aligned}
\alpha(i) &= a + v^T(I-A)^{-1}w \\
&= 1 - \epsilon - v^T\mathbb{1} + v^T(I-A)^{-1}w \\
&= 1 - \epsilon - \left((I-A)^T\mathbb{1} - \epsilon\mathbb{1}\right)^T\mathbb{1} \\
&\quad + \left((I-A)^T\mathbb{1} - \epsilon\mathbb{1}\right)^T(I-A)^{-1}\left((I-A)\mathbb{1} - \epsilon\mathbb{1}\right) \\
&= 1 - \epsilon - \mathbb{1}^T(I-A)\mathbb{1} + \epsilon\mathbb{1}^T\mathbb{1} \\
&\quad + \mathbb{1}^T(I-A)\mathbb{1} - 2\epsilon\mathbb{1}^T\mathbb{1} + \epsilon^2\mathbb{1}^T(I-A)^{-1}\mathbb{1} \\
&= 1 - n\epsilon + \epsilon^2\mathbb{1}^T(I-A)^{-1}\mathbb{1}.
\end{aligned}$$

So, to show that $\alpha(i) \geq (1-\epsilon)^n$, we will show that

$$\mathbb{1}^T(I-A)^{-1}\mathbb{1} \geq \frac{(1-\epsilon)^n - (1-n\epsilon)}{\epsilon^2},$$

with equality if and only if

$$C = \frac{1}{1-\epsilon}B$$

is a cyclic permutation matrix. By Lemma B.11,

$$\sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k = \frac{(1-\epsilon)^n - (1-n\epsilon)}{\epsilon^2}.$$

So, we need to show that

$$\mathbb{1}^T(I-A)^{-1}\mathbb{1} \geq \sum_{k=0}^{n-2}(n-k-1)(1-\epsilon)^k.$$

Via Lemma 2.8, we have

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

By Lemma B.10, $\mathbb{1}^T A^k \mathbb{1} \geq (n-k-1)(1-\epsilon)^k$ whenever $0 \leq k \leq n-2$. Therefore,

$$\begin{aligned}
\mathbb{1}^T(I-A)^{-1}\mathbb{1} &= \mathbb{1}^T \left( \sum_{k=0}^{\infty} A^k \right) \mathbb{1} \\
&= \sum_{k=0}^{n-2} \mathbb{1}^T A^k \mathbb{1} + \mathbb{1}^T \left( \sum_{k=n-1}^{\infty} A^k \right) \mathbb{1} \\
&\geq \sum_{k=0}^{n-2} \mathbb{1}^T A^k \mathbb{1} \\
&\geq \sum_{k=0}^{n-2} (n-k-1)(1-\epsilon)^k.
\end{aligned}$$

We note that equality occurs if and only if $\mathbb{1}^T A^k \mathbb{1} = (n-k-1)(1-\epsilon)^k$ for $0 \leq k \leq n-1$ and

$$\sum_{k=n-1}^{\infty} A^k = 0;$$

the second condition is equivalent to $A^{n-1} = 0$ (since $A$ is nonnegative).

Thus,

$$\mathbb{1}^T(I-A)^{-1}\mathbb{1} \geq \frac{(1-\epsilon)^n - (1-n\epsilon)}{\epsilon^2}$$

and we see that

$$\begin{aligned}
\alpha(1) &= a + v^T(I-A)^{-1}w \\
&= 1 - n\epsilon + \epsilon^2 \mathbb{1}^T(I-A)^{-1}\mathbb{1} \\
&\geq 1 - n\epsilon + \epsilon^2 \tfrac{(1-\epsilon)^n - (1-n\epsilon)}{\epsilon^2} \\
&= (1-\epsilon)^n.
\end{aligned}$$

Suppose that we have equality; that is, suppose that $\alpha(1) = (1-\epsilon)^n$. As noted above, this occurs if and only if $\mathbb{1}^T A^k \mathbb{1} = (n-k-1)(1-\epsilon)^k$ for $0 \leq k \leq n-2$ and $A^{n-1} = 0$. When the matrix $A$ satisfies $A^{n-1} = 0$, it is nilpotent. It is well-known (see [3]) that a nonnegative matrix is nilpotent if and only if it is permutation-similar to an upper-triangular matrix. That is, since $A^{n-1} = 0$, there is a permutation matrix $P$ (of order $n-1$) such that

$$PAP^T = \begin{bmatrix} 0 & * & \cdots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ & & & 0 \end{bmatrix},$$

where the entries below the diagonal are zeroes. Now, we label the entries in the first diagonal as $a_1, \ldots, a_{n-2}$ (the matrix $A$ has order $n-1$). That is, let

$$PAP^T = \begin{bmatrix} 0 & a_1 & * & \cdots & * \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & * \\ & & & \ddots & a_{n-1} \\ & & & & 0 \end{bmatrix}.$$

Then,

$$PA^{n-2}P^T = \begin{bmatrix} 0 & \cdots & 0 & \prod_{k=1}^{n-2} a_k \\ & \ddots & & 0 \\ & & \ddots & \vdots \\ & & & 0 \end{bmatrix}.$$

We have $\mathbb{1}A^{n-2}\mathbb{1} = (n-(n-2)-1)(1-\epsilon)^{n-2} = (1-\epsilon)^{n-2}$, implying that

$$\prod_{k=1}^{n-2} a_k = (1-\epsilon)^{n-2}.$$

Since the matrix $B$ is equal to a stochastic matrix multiplied by the scalar $1-\epsilon$, we have $a_k \leq 1-\epsilon$ for all $k$. This, together with the above equality, implies that, in fact, $a_k = 1-\epsilon$ for all $k$. Then, we also have

$$\mathbb{1}^T A \mathbb{1} = (n-1-1)(1-\epsilon) = (n-2)\epsilon.$$

Since the terms $a_k$ are each equal to $1-\epsilon$ and there are $n-2$ of them, the remainder of the entries in $A$ must be 0. So, in fact,

$$PAP^T = \begin{bmatrix} 0 & 1-\epsilon & & \\ & \ddots & \ddots & \\ & & \ddots & 1-\epsilon \\ & & & 0 \end{bmatrix},$$

where the unspecified entries are zeroes. Now, we have $A\mathbb{1} + w = (1-\epsilon)\mathbb{1}$ and $A^T\mathbb{1} + v = (1-\epsilon)\mathbb{1}$. So, since $P\mathbb{1} = P^T\mathbb{1} = \mathbb{1}$, we have

$$Pw = (1-\epsilon)\mathbb{1} - PAP^T\mathbb{1} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1-\epsilon \end{bmatrix}$$

$$\text{and } Pv = (1-\epsilon)\mathbb{1} - PA^TP^T\mathbb{1} = \begin{bmatrix} 1-\epsilon \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

We also have $a = 1 - \epsilon - v^T\mathbb{1} = 0$ (from the above conclusion concerning $v$). Thus, since

$$B = \begin{bmatrix} a & v^T \\ w & A \end{bmatrix},$$

we see that if $\alpha(1) = (1-\epsilon)^n$, then there is a permutation matrix $P$ of order $n-1$ such that

$$\begin{bmatrix} 1 & 0 \\ 0 & P \end{bmatrix} B \begin{bmatrix} 1 & 0 \\ 0 & P^T \end{bmatrix} = \begin{bmatrix} 0 & v^TP^T \\ Pw & PAP^T \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1-\epsilon & & \\ & \ddots & \ddots & \\ & & \ddots & 1-\epsilon \\ 1-\epsilon & & & 0 \end{bmatrix},$$

where the unspecified entries are zeroes. Via Proposition B.9, we see that if $B = (1-\epsilon)C$ where $C$ is a cyclic permutation matrix, then for all $i \in \mathcal{E}$, we have $\alpha(i) = (1-\epsilon)^n$.

∎

# Appendix C

# Data analysis

## C.1 *n*-Pentane analysis

We first examine two data sets obtained from experiments concerning the $n$-pentane molecule, $CH_3 - (CH_2)_3 - CH_3$. The data is obtained from two hybrid Markov chain Monte Carlo experiments, using temperature parameters of 300 and 500.

We first summarise, briefly, the concept of a hybrid Markov chain Monte Carlo experiment (HMCMC). Suppose that $X = \{x_t\}$ is a Markov chain on the state space $\mathcal{S} = \mathbb{R}^n$ or $\mathbb{C}^n$. An HMCMC experiment is a manner in which a second Markov chain $Y = \{y_t\}$ may be realised, via mathematical software, which models or simulates $X$. The transition probabilities of $X$ are not utilised in this simulation, and so HMCMC experiments are useful if these probabilities are unknown or difficult to calculate. The only inputs required are the stationary distribution $\pi$ of $X$ and a temperature

parameter $T$.

In a Markov chain on a continuous state space, the stationary distribution $\pi$ is a probability measure. That is, there is a function $\mu : \mathcal{S} \to [0, 1]$ such that for any $\mathcal{E} \subseteq \mathcal{S}$,

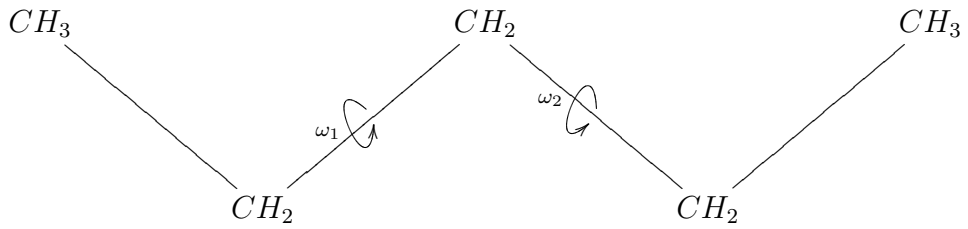
$$\pi(\mathcal{E}) = \int_{\mathcal{E}} \mu(x)dx.$$

The HMCMC method assumes that the Markov chain $X$ is a discretisation, time-wise, of a process which is continuous in time. That is, it is assumed that if $x_0 = i_0, x_1 = i_1, \ldots$ is a realisation of $X$, then there is a continuous function $f : [0, \infty) \to \mathcal{S}$ such that $f(t) = x_t$ whenever $t$ is an integer. The process $Y$ is based on Hamiltonian mechanics – it simulates a particle moving through the state space $\mathcal{S}$ with a velocity which randomly fluctuates, subject to the constraint that for any $\mathcal{E} \subseteq \mathcal{S}$, $\pi(\mathcal{E})$ is approximately equal to the frequency with which $Y$ visits states in $\mathcal{E}$. Thus, although the individual stepwise transitions $x_t \to x_{t+1}$ and $y_t \to y_{t+1}$ may not follow the same rules, they are both discretisations of continuous processes which have the same stationary distribution. Thus, $Y$ is seen to be a useful model of $X$.

The temperature parameter $T$ input into a HMCMC experiment controls the volatility of the velocity of $Y$. When $T$ is small, one tends to see $y_{t+1} - y_t$ remain fairly constant, for long periods of time, and $\|y_{t+1} - y_t\|$ remain bounded by a small value, overall. For large $T$, the difference vector $y_{t+1} - y_t$ can change more rapidly (as a function of $t$) and become larger in norm.

In a HMCMC experiment (in fact, in all Monte Carlo methods) the simulation $Y$ is a reversible Markov chain. An introduction to Markov chain Monte Carlo, in general, appears in [21, Section 5.5]; a detailed explanation of the experiment used to model the $n$-pentane molecule appears in [22].

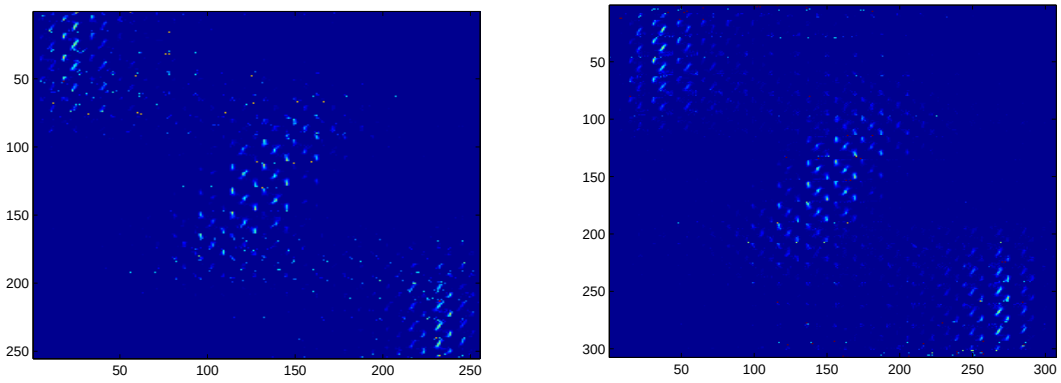
The state of the $n$-pentane molecule is determined by the two dihedral angles between the $CH_3 - CH_2$ components of the molecule and the remaining $CH_2$ component.




The range of attainable angles is discretised into 20 intervals, creating a state space $\mathcal{S}$ of order $20^2 = 400$ which is isomorphic to a subset of $\mathbb{R}^2$. As time proceeds, these angles change randomly and it is assumed this process satisfies the Markov Property. The stationary distribution is known, based on the molecular properties of $CH_3 - (CH_2)_3 - CH_3$. So, the HMCMC method is used to construct a simulation of the random changes in the molecular states of the $n$-pentane molecule. (See [7, 22] for details.)

The authors of [7] construct two distinct sequences

$$y_0^{(1)}, y_1^{(1)}, \ldots, y_{s_1}^{(1)} \text{ and } y_0^{(2)}, y_1^{(2)}, \ldots, y_{s_2}^{(2)}$$




(a) $Ph300$ (b) $Ph500$

Figure C.1: The $n$-pentane transition matrices

using a hybrid Markov chain Monte Carlo algorithm – the first corresponding to temperature $T = 300$ and the second to $T = 500$. These sequences are then used to construct two transition matrices $Ph300 = A^{(1)}$ and $Ph500 = A^{(2)}$, defined via

$$a_{ij}^{(k)} = \frac{\left|\{t \leq s_k - 1 : y_t^{(k)} = i \text{ and } y_{t+1}^{(k)} = j\}\right|}{\left|\{t \leq s_k - 1 : y_t^{(k)} = i\}\right|}.$$

A graphic representation of these matrices appears in Figure C.1; lighter pixels represent significant entries and darker pixels represent entries near 0. Neither experiment results in all 400 potential states being observed. In the case of temperature 300, 255 distinct states appear and in the case of temperature 500, 307 states are observed. The stationary distributions $\pi^{(1)}$ and $\pi^{(2)}$ of these matrices are provided by the authors. It can be confirmed, using the data provided, that for $k = 1$ and 2,

$$\sum_i \sum_j \left|\pi_j^{(k)} a_{ij}^{(k)} - \pi_i^{(k)} a_{ij}^{(k)}\right|$$

| Algorithm | $\epsilon$ | Agg. | $\pi$-Min. | $\mathbb{1}$-Min. | Refined $\mathbb{1}$-min. | Near Tr. States |
|---|---|---|---|---|---|---|
| MMaxE | 0.01 | 7 | 0.918 | 0.803 | 0.836 | 163 |
| | 0.005 | 5 | 0.979 | 0.823 | 0.907 | 135 |
| LW | 0.01 | 7 | 0.921 | 0.768 | 0.835 | 159 |
| | 0.005 | 5 | 0.979 | 0.768 | 0.910 | 116 |
| PO | 0.01 | 7 | 0.920 | 0.768 | 0.852 | 158 |
| | 0.005 | 5 | 0.979 | 0.768 | 0.911 | 117 |
| MinC | 0.01 | 7 | 0.919 | 0.770 | 0.835 | 167 |
| | 0.005 | 5 | 0.978 | 0.770 | 0.907 | 134 |
| PCCA | n/a | 7 | 0.918 | n/a | n/a | n/a |
| | n/a | 5 | 0.976 | n/a | n/a | n/a |
| SVD | n/a | 7 | 0.876 | 0.659 | n/a | n/a |

Table C.1: Stochastic complement based and other algorithms applied to $Ph300$

is equal to a near-negligible positive number; so the matrices $Ph300$ and $Ph500$ are reversible with known stationary distributions. The reversible property and the known stationary distributions allow us to apply every one of our stochastic complement based algorithms to $Ph300$ and $Ph500$; as well, we are able to utilise the $\pi$-coupling measure in evaluating our output.

We first examine the $255 \times 255$ matrix $Ph300$. The eight eigenvalues of $Ph300$ with largest magnitude are

$$\{1, 0.986, 0.984, 0.982, 0.975, 0.941, 0.938, 0.599\}.$$

In [7], the PCCA Algorithm is applied to $Ph300$ twice, once to decouple the state space into 5 aggregates and once to decouple the state space into 7 aggregates. The matrix $Ph300$ has 4 eigenvalues that are approximately 0.98 and a further 2 that are approximately 0.94 – when combined with the eigenvalue 1, this suggests a Perron

cluster of either 5 or 7 eigenvalues.

The authors of [7] use the $\pi$-coupling measure to evaluate their obtained aggregates. Given a stochastic matrix $A$ on the state space $\mathcal{S}$ with stationary distribution $\pi$, the $\pi$-coupling measure of a constructed aggregate $\mathcal{E} \subseteq \mathcal{S}$ is the value

$$w_\pi(\mathcal{E}) = \frac{\pi(\mathcal{E})^T A(\mathcal{E}, \mathcal{E})\mathbb{1}}{\pi(\mathcal{E})^T \mathbb{1}} = \frac{\sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{E}} \pi_i a_{ij}}{\sum_{i\in\mathcal{E}} \pi_i}.$$

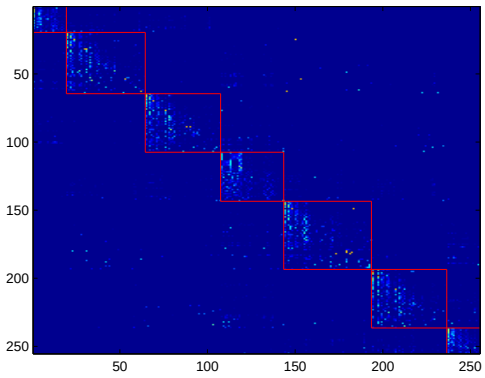
The authors of [7] evaluate a potential decoupling $\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m)$ of the state space via the $\pi$-coupling measure. If the minimum of the values $\{w_\pi(\mathcal{E}_k)\}$ is close to 1, then $\Psi$ is seen to be a good uncoupling. We will follow this convention, and produce the minimum coupling measure of the outputs of our algorithms.

The weakest $\pi$-coupling measure of an aggregate of $Ph300$ obtained by the PCCA Algorithm is 0.976 in the case of 5 aggregates and is 0.918 in the case of 7 aggregates (see [7] for a full analysis of the algorithm's performance).

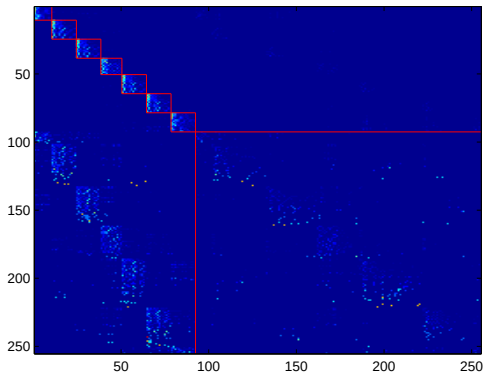
In [10] the authors apply the SVD-based algorithm (Algorithm 2) to the matrix $Ph300$. As with the PCCA approach, 7 aggregates are obtained. However, the coupling measures of the obtained aggregates are somewhat lower. The minimum $\pi$-coupling measure of an aggregate obtained is 0.876. As well, the authors examine the $\mathbb{1}$-coupling measure,

$$w_{\mathbb{1}}(\mathcal{E}) = \frac{\sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{E}} a_{ij}}{|\mathcal{E}|};$$

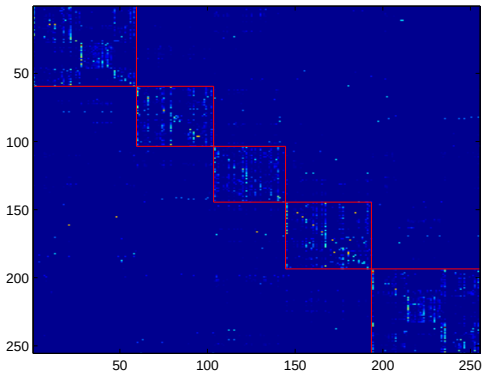
the minimal $\mathbb{1}$-coupling measure is 0.659.



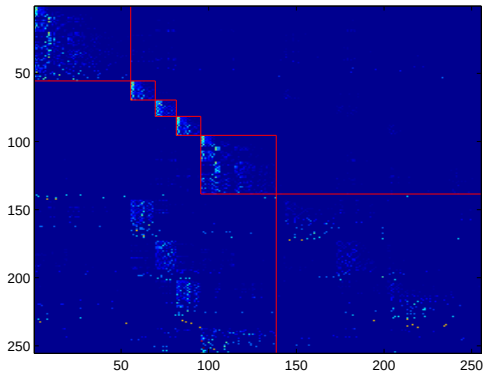

(a) MMaxE with $\epsilon = 0.01$,




(b) followed by the Refinement Algorithm.




(c) PO with $\epsilon = 0.005$,




(d) followed by the Refinement Algorithm.

Figure C.2: Modified Maximum Entry (MMaxE) and Perron Ordered (PO) Algorithms applied to $Ph300$, followed by the Refinement Algorithm.

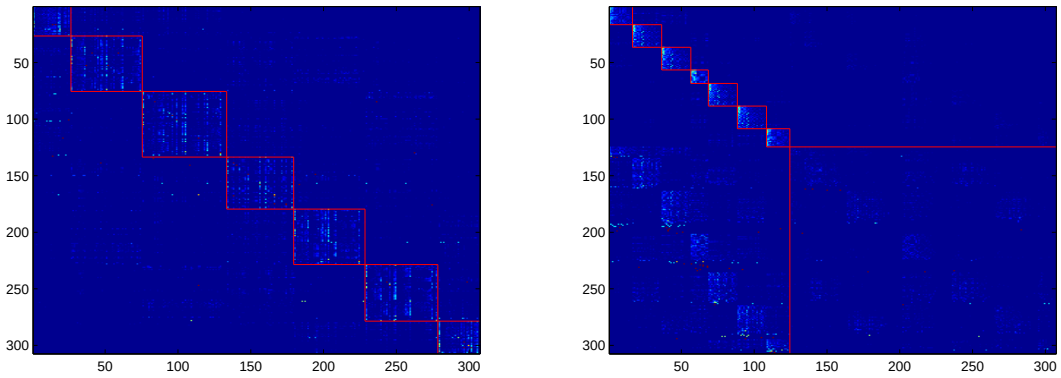
We apply the Modified Maximum Entry (MMaxE), Lower-weighted (LW), Perron Ordered (PO) and Minimum Column (MinC) Algorithms with inputs $\epsilon = 0.01$ and 0.005. (Experiments using a range of values for $\epsilon$ show that these values produce 7 and 5 aggregates, respectively). As the matrix $Ph300$ is reversible, we use, in all cases, the criteria

$$\hat{a}_{ii} < \frac{(1-\epsilon)^2}{1+(m_i-2)\epsilon}$$

to determine if the state $i$ is safe to remove via a stochastic complement (where $\hat{A}$ is the stochastic complement under consideration during some iteration of one of our algorithms). In Table C.1, we show the smallest $\pi$-coupling measure and the smallest $\mathbb{1}$-coupling measure of an obtained aggregate for each application (of our algorithms and others').

In addition, after applying each of these four algorithms, we apply the Refinement Algorithm in an attempt to identify a collection of near-transient states. We show the smallest $\mathbb{1}$-coupling measure of an almost invariant aggregate after the near transient states (identified by the Refinement Algorithm) have been removed from each aggregate, along with the total number of near transient states the algorithm identified.

As the $\pi$-coupling measure significantly reduces the contribution of near transient states, the $\pi$-coupling measures of the aggregates tend to be altered only insignificantly after applying the Refinement Algorithm; thus, we do not produce the




(a) LW with $\epsilon = 0.05$, (b) followed by the Refinement Algorithm.

Figure C.3: Lower-weighted Algorithm (LW) and Refinement Algorithms applied to $Ph500$

$\pi$-coupling measures of the refined aggregates.

It is interesting to note that this matrix seems to have a particularly large number of near transient states (as we noted above, the total order of the matrix is 255). The outputs for every one of our algorithms compare very favourably to those obtained in [7] and [10].

We present graphical representations of four of our outputs in Figure C.2. In Figures C.2b and C.2d, the final diagonal block does not represent an almost invariant aggregate – it corresponds to the collection of states identified by the Refinement Algorithm as near-transient. The outputs of the modified Maximum Entry and Perron ordered Algorithms, depicted in Figures C.2a and C.2c, clearly show almost invariant aggregates of the Markov chain. However, we see many significant entries not contained in the diagonal blocks. After applying the Refinement Algorithm, the structure

| Algorithm | $\epsilon$ | Agg. | $\pi$-Min. | $\mathbb{1}$-Min. | Refined $\mathbb{1}$-min. | Near Tr. States |
|---|---|---|---|---|---|---|
| MMaxE | 0.05 | 7 | 0.770 | 0.659 | 0.712 | 198 |
| | 0.01 | 5 | 0.881 | 0.694 | 0.782 | 172 |
| LW | 0.05 | 7 | 0.786 | 0.659 | 0.718 | 183 |
| | 0.01 | 5 | 0.881 | 0.659 | 0.781 | 170 |
| PO | 0.05 | 7 | 0.786 | 0.659 | 0.718 | 183 |
| | 0.01 | 5 | 0.881 | 0.659 | 0.781 | 169 |
| MinC | 0.05 | 7 | 0.787 | 0.657 | 0.718 | 151 |
| | 0.01 | 5 | 0.881 | 0.675 | 0.786 | 168 |

(a) Stochastic complement based algorithms applied to $Ph500$

| Partition vector | Termination Criteria | Agg. | $\pi$-Min. | $\mathbb{1}$-Min. |
|---|---|---|---|---|
| Left-singular | $\mathbb{1}$-coupling | 5 | 0.584 | 0.552 |
| Left-singular | $\pi$-coupling | 6 | 0.584 | 0.490 |
| Right-singular | $\mathbb{1}$-coupling | 6 | 0.656 | 0.507 |

(b) SVD based algorithm applied to $Ph500$

Table C.2: Stochastic complement and SVD based algorithms applied to $Ph500$

described in Lemma 3.5 is more apparent.

We next turn to the matrix $Ph500$ (Figure C.1b). This matrix does not appear in [7] but is examined in [10].

As stated in the description of the SVD-based algorithm (Algorithm 2), there are two different choices for each of of two methods within its implementation, resulting in four different possible implementations. The user may utilise either right or left-singular vectors and either of the $\pi$ or $\mathbb{1}$-coupling measures. In the application of the SVD-based algorithm to $Ph300$, left-singular vectors were used and it was noted that using either coupling measure produced the same output. In the analysis of $Ph500$, the SVD-based algorithm is applied three times – the implementation which

uses right-singular vectors and the $\pi$-coupling measure is not applied.

In Table C.2, we examine the outputs of the Modified Maximum Entry, Lower-weighted, Perron Ordered and Minimum Column Algorithms, in addition to the SVD based algorithm. We use the input values $\epsilon = 0.05$ and 0.01. As above, we show the number of aggregates produced, the minimal coupling measures of an aggregate, the minimal $\mathbb{1}$-coupling measure of an aggregate after applying the Refinement Algorithm and the number of near-Transient states identified by the Refinement Algorithm. We note that, as above, the Refinement Algorithm has identified a somewhat large number of near transient states (the total number of states is 307).

## C.2 A collaboration network

We discuss a particularly interesting, somewhat problematic example of a nearly uncoupled Markov chain. This example is taken from [20], where it is used to analyse *network centrality* – how "important" a given vertex is within a network. The example is a collaboration network, given in the form of a weighted graph; each node represents a researcher (working, specifically, on network-related research) and the (undirected) edges represent collaborations (papers co-authored) between researchers.

Suppose that researchers $i$ and $j$ have co-authored $k$ papers; let $n_1, \ldots, n_k$ be the total number of authors of each of these $k$ papers, respectively. Then, the weight of the $ij$th edge is

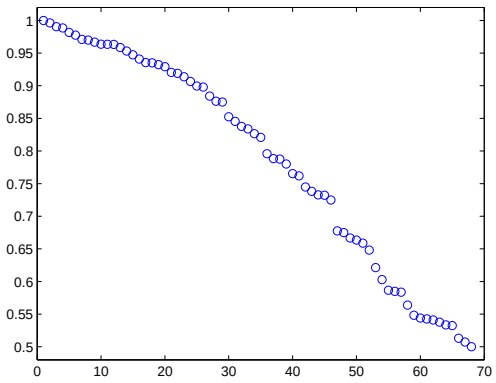



Figure C.4: The 69 eigenvalues of $A^{(c)}$ closest to 1

$$w_{ij} = \sum_{l=1}^{k} \binom{n_l}{2}.$$

The weights are chosen so that each paper contributes a total weight of 1 to the edges of the graph. (Papers with a single author are not considered.)

There are 1589 researchers in total; however, the graph is not connected. We examine (as in [20]) the largest connected component only, which contains 379 vertices.

We will apply the Maximum Entry Algorithm to the random walk on this weighted graph. So, we examine the $379 \times 379$ reversible stochastic matrix $A^{(c)}$ with $ij$th entry defined via

$$a_{ij}^{(c)} = \frac{w_{ij}}{\sum_{k=1}^{379} w_{ik}},$$

where $w_{ij}$ is as defined above.

| $\delta$ | Agg. | $\pi$-Min. | $\mathbb{1}$-Min. | Average $\mathbb{1}$-coupling Measure |
|---|---|---|---|---|
| 0.20 | 27 | 0.745 | 0.756 | 0.930 |
| 0.15 | 18 | 0.745 | 0.756 | 0.943 |
| 0.10 | 10 | 0.867 | 0.905 | 0.960 |
| 0.05 | 2 | 0.993 | 0.992 | 0.996 |

Table C.3: The Maximum Entry Algorithm applied to the collaboration matrix $A^{(c)}$

Perhaps unsurprisingly, the random walk on this largest connected component is very much nearly uncoupled. However, its eigenvalues do not clearly identify the number of potential aggregates. The matrix $A^{(c)}$ has 69 eigenvalues between 0.5 and 1, and they seem to be evenly distributed within this interval (Figure C.4).

First, we apply the Maximum Entry Algorithm to $A^{(c)}$ with inputs $\delta = 0.20$, 0.15, 0.10 and 0.05. We summarise the results of these applications in Table C.3. As well, we include the average $\mathbb{1}$-coupling measure of the aggregates obtained. The Refinement Algorithm identifies only a tiny number of near-transient states (five or fewer for each output) and so we do not include its output.

The average $\mathbb{1}$-coupling measures of the produced aggregates are quite high - significantly higher than the the minimum measures. Closer examination shows that most of the obtained aggregates have $\mathbb{1}$-coupling measures that are very close to these averages; *i.e* there are only a small number of outliers that have significantly lower coupling measures.

The output digraph is not necessary for every application – it is straightforward to modify our algorithms so that only the aggregates themselves are output. However,

| $\alpha$ | Agg. | Average $\mathbb{1}$-coupling Measure |
|---|---|---|
| 0.85 | 27 | 0.938 |
| 0.90 | 23 | 0.947 |
| 0.95 | 12 | 0.971 |

Table C.4: Recursive subaggregating applied to the collaboration network

the digraph is a great source of information on the uncoupled structure of the matrix. We will show one example of how it can be used to produce a hierarchical uncoupling of the matrix.

We will first show how, given $\alpha < 1$, we can produce a partition $\Psi$ such that for every $\mathcal{E} \in \Psi$, $w_{\mathbb{1}}(\mathcal{E}) > \alpha$ (eliminating the possibility we saw above of outlying aggregates with smaller coupling measures).

We construct the digraph $G$ by applying the Maximum Entry Algorithm to $A = A^{(c)}$ with $\delta = 0$. As well, we record the order in which the directed arcs were added to $G$. Thus, $G$ is weakly connected (since $A$ is irreducible), acyclic and every vertex has out-degree equal to 0 or 1. The fact that $G$ is weakly connected implies that there is a unique vertex with out-degree 0.

We note that removing $k$ directed arcs from $G$ results in a digraph that is acyclic and has every out-degree equal to 0 or 1; as well, the number of vertices with out-degree 0 is increased by exactly $k$. Thus, by Lemma 5.2, removing $k$ directed arcs from $G$ results in an acyclic digraph $G'$ where every vertex has out-degree equal to 0 or 1 which contains exactly $k + 1$ weakly connected components.

We apply the following iterative procedure to $G$.

1. Let $G' = G$.

2. Let $G_0$ be a weakly connected component of $G'$ (during the first iteration, $G_0 = G' = G$) and let $V_0$ be the vertices of $G'$.

3. Let $i \to j$ be the final directed arc with $i, j \in V_0$ added to $G$ by the algorithm.

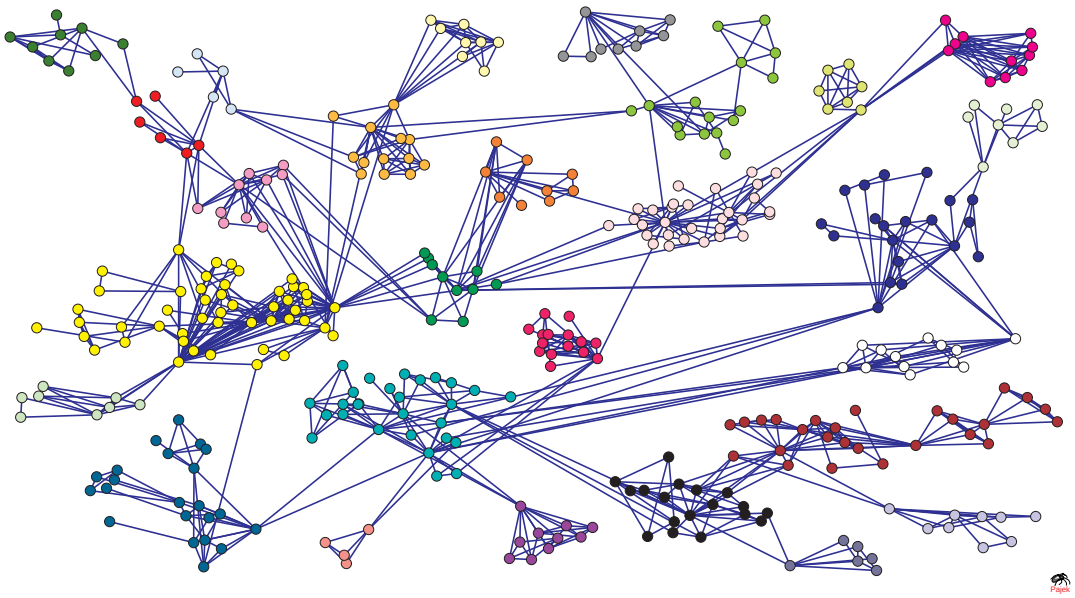
4. Let $G_1$ and $G_2$ be the weakly connected subgraphs of $G_0$ that are obtained by removing the directed arc $i \to j$ and let $V_1$ and $V_2$ be their respective vertex sets.

5. If the $\mathbb{1}$-coupling measures of $V_1$ and $V_2$ are both strictly greater than $\alpha$, remove the directed arc $i \to j$ from $G'$.

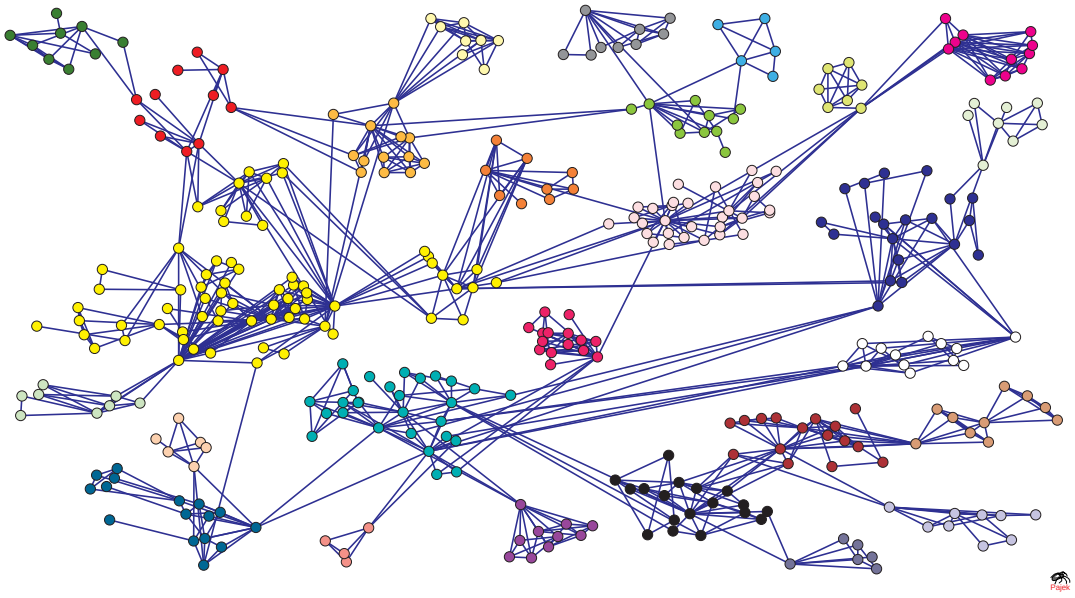
6. We repeat steps 2 through 5 until every weakly connected component $G_0$ of $G$ is such that removing the final arc with endpoints in $V_0$ added by the Maximum Entry Algorithm results in at least one of $V_1$ or $V_2$ having $\mathbb{1}$-coupling measure less than or equal to $\alpha$.

We refer to this process as recursive subaggregating. The directed arcs that the Maximum Entry Algorithm adds first are those that we are most sure represent transitions within an almost invariant aggregate – they represent maximal transition probabilities within stochastic complements where very few states have been removed.

Thus, after running the Maximum Entry Algorithm with $\delta = 0$, the directed arc added last seems most likely to be a transition from one aggregate to another. If the



(a) The Maximum Entry Algorithm, with $\delta = 0.2$.



(b) Recursive Subaggregating, with $\alpha = 0.85$.

Figure C.5: Both techniques, applied to the collaboration matrix $A^{(c)}$, produce 27 aggregates.

two weakly connected components obtained by removing this arc are, indeed, almost invariant aggregates, we remove this arc and then apply the same reasoning to each of the two weakly connected components.

This process is guaranteed to produce a collection $\Psi$ of almost invariant aggregates that each have $\mathbb{1}$-coupling measure greater than $\alpha$. Although, it is entirely possible that recursive subaggregation may return very few (or even just one) such aggregates.

Then, we may select a value $\alpha_2$ such that $\alpha_2 < \alpha$ and apply recursive subaggregation with the value $\alpha_2$ to the digraph $G'$. This returns a partition $\Psi_2$ of $\mathcal{S}$ such that

1. for every $\mathcal{E}' \in \Psi_2$, $w_{\mathbb{1}}(\mathcal{E}') > \alpha_2$, and

2. for every $\mathcal{E}' \in \Psi_2$, there is $\mathcal{E} \in \Psi$ such that $\mathcal{E}' \subseteq \mathcal{E}$.

This process may be repeated any number of times, producing a hierarchy of almost invariant aggregates of $A$.

We apply this procedure three times to the digraph $G$ obtained from the collaboration network described above (via the Maximum Entry Algorithm with $\delta = 0$). We present our results in Table C.4. We emphasise that every aggregate obtained has $\mathbb{1}$-coupling measure greater than the input $\alpha$. For example, recursive subaggregating with $\alpha = 0.85$ returns 27 collections of states which each have a $\mathbb{1}$-coupling measure greater than 0.85 and have an average $\mathbb{1}$-coupling measure of 0.938. This is a significant improvement of the aggregates obtained from the Maximum Entry Algorithm

with $\delta = 0.2$ (which each have a $\mathbb{1}$-coupling measure greater than 0.75 and have an average $\mathbb{1}$-coupling measure of 0.930). We show these two outputs in Figures C.5a and C.5b.

This is just one example of how the output digraph may be utilised. The transition digraph of a stochastic matrix order $n$ can contain on the order of $n^2$ directed arcs; it can be computationally difficult to extract any meaningful information from such a structure. However, the output digraph of our stochastic complement based algorithms contains approximately $n$ directed arcs, and these arcs can be ranked in significance by the order in which they were added to the digraph. This is a very rich source of information on the probabilistic properties of the associated state space.

## C.3 Randomly generated examples

We present a summary of the Lower Weighted Algorithm's performance when applied to a collection of randomly generated reversible stochastic matrices.

We require a method of generating a random unweighted graph such that the associated random walk is nearly uncoupled with respect to $\epsilon$. For our purposes, the output of Algorithm 12, seems sufficient. The inputs of Algorithm 12 are positive integers $m$ and $n$ with $2 < m < n$ and a positive value $\epsilon < 1$. The output is a graph on vertices $V = \{1, \ldots, n\}$ and a partition $(\mathcal{E}_1, \ldots, \mathcal{E}_m)$ of the vertex set. We show below that if $n/m$ is sufficiently large, each $\mathcal{E}_k$ is an almost invariant aggregate of the

random walk on the output graph.

**Algorithm 12** Random graph generator

Let $G$ be the graph on vertices $V = \{1, \ldots, m\}$ that contains no edges.
**for** $k = 1, \ldots, m$ **do**
  $\mathcal{E}_k := \{k\}$
**end for**
**for** $i = m + 1, \ldots, n$ **do**
  Choose $k \in \{1, \ldots, m\}$, uniformly.
  Choose $p \in (0, 1)$, uniformly.
  Let $q = \min \left\{ 1, \frac{\epsilon p |\mathcal{E}_k|}{(1-\epsilon) \sum_{l \neq k} |\mathcal{E}_l|} \right\}$.
  Add vertex $i$ to $V$.
  For each $j \in \mathcal{E}_k$, the edge $ij$ is added to $G$ with probability $p$.
  For each $j \in \bigcup_{l \neq k} \mathcal{E}_l$, the edge $ij$ is added to $G$ with probability $q$.
  Add vertex $i$ to $\mathcal{E}_k$.
**end for**
**return** $G$ and $(\mathcal{E}_1, \ldots, \mathcal{E}_m)$.

The random choices in Algorithm 12 are assumed to have uniform distributions – when an element is chosen from a set, every member of that set is equally likely, when a value $p \in (0, 1)$ is chosen, the expected value of $p$ is $1/2$, and so forth.

If, after some number of iterations, $\mathcal{E}_k$ contains $r$ vertices, and we then add one more vertex $i$ to $\mathcal{E}_k$, the expected number of new edges $ij$ with $j \in \mathcal{E}_k$ is $r\mathbb{E}[p] = r/2$. Thus, after adding $s$ vertices to $\mathcal{E}_k$, the expected number of edges in the induced subgraph $G(\mathcal{E}_k)$ is

$$\sum_{r=1}^{s-1} \frac{r}{2} = \frac{s(s-1)}{4} = \frac{1}{2} \binom{s}{2}.$$

The complete graph on $s$ vertices (the graph which contains every possible edge) contains $\binom{s}{2}$ edges. If the choice of $k \in \{1, \ldots, m\}$ within the algorithm is accomplished so that each possibility is equally likely, the expected size of each aggregate $\mathcal{E}_k$ is $n/m$. So, as long as the number $n/m$ is somewhat large, the $m$ induced subgraphs $G(\mathcal{E}_k)$ each contain approximately half of the maximum possible number of edges. This, together with the fact that the edges are added in a random manner, implies that (again, if $n/m$ is large) each induced subgraph $G(\mathcal{E}_k)$ is well connected.

Now, suppose that at some iteration, we are adding a vertex $i$ to the aggregate $\mathcal{E}_k$ and that the probability $p$ of connecting $i$ to the members of $\mathcal{E}_k$ has already been selected. Let

$$q = \min \left\{ 1, \frac{\epsilon p \left| \mathcal{E}_k \right|}{(1-\epsilon) \sum_{l \neq k} \left| \mathcal{E}_k \right|} \right\},$$

let

$$a = p \left| \mathcal{E}_k \right|$$

be the expected number of edges $ij$ with $j \in \mathcal{E}_k$ added at this iteration and let

$$b = q \sum_{l \neq k} \left| \mathcal{E}_l \right| \leq \frac{\epsilon p \left| \mathcal{E}_k \right|}{1-\epsilon}$$

be the expected number number of new edges $ij$ with $j \notin \mathcal{E}_k$ added. Then,

$$\frac{b}{a+b} \leq \frac{\epsilon p \left| \mathcal{E}_k \right| / (1-\epsilon)}{p \left| \mathcal{E}_k \right| + \epsilon p \left| \mathcal{E}_k \right| / (1-\epsilon)} = \frac{\epsilon/(1-\epsilon)}{1+\epsilon/(1-\epsilon)} = \epsilon.$$

Thus, the expected value of the ratio of the number of edges with endpoints in different aggregates to the total number of edges in $G$ is less than or equal to $\epsilon$. So, as long as $n$ is large, the probability that a randomly selected edge has endpoints in distinct aggregates is approximately bounded by $\epsilon$. Thus, for large values of $n$ and $n/m$, the random graph generator proposed above produces a graph whose random walk is nearly uncoupled with respect to $\epsilon' \approx \epsilon$.

We have specified that $p$, with the algorithm, be chosen randomly so that there is some variation in the degrees of the vertices of $G$.

As an example of our above discussion we have constructed a graph $G$ where one of the aggregates $\mathcal{E}_k$ contains 50 vertices. We calculate the stochastic matrix $A$ corresponding to the random walk on this induced subgraph $G(\mathcal{E}_k)$; the eigenvalue of $A$ closest, but not equal, to 1 is $\lambda \approx 0.3093$. So, this subgraph is very well-connected (see Proposition 3.7). The degrees of the vertices in the induced subgraph $G(\mathcal{E}_k)$ are given in Figure C.6.

We apply our Lower Weighted Algorithm to matrices generated by Algorithm 12. We generate a total of 180 random graphs, calculate the stochastic matrices of their associated random walks and then apply Algorithm 8. Every graph generated has $n = 1000$ vertices.

In a sense, there is a danger to using randomly generated matrices to test algorithms such as ours. It is somewhat easy to "fine-tune" the inputs so that the

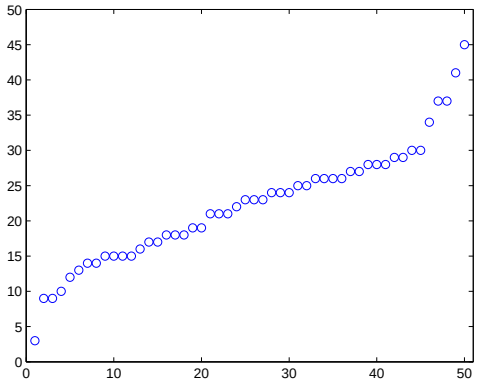



Figure C.6: The vertex degrees within an aggregate output by Algorithm 12.

algorithm will succeed. For example, if the number of almost invariant aggregates is known, one can force an algorithm to output that exact number of aggregates. So, as much as possible, we try to apply the Lower Weighted Algorithm without taking advantage of the known structure of the matrices we generate. In particular, if the algorithm outputs the wrong number of aggregates, we will simply report the coupling measures of this output without attempting to refine or alter it. We emphasise that this is not how the algorithms we have presented should be used in practise – the speed of the algorithms allows many applications to a single data set in a relatively short period of time, so that a particularly optimal output may be chosen or constructed.

In applying our stochastic complement based algorithms, we have found that the value $\delta = 0.05$ tends to be a very good first input. If the matrix is genuinely nearly uncoupled, this fact seems to be apparent in examining the aggregates output with

this first input value. The cheapness (in speed terms) of the algorithm then allows us to fine-tune the input $\delta$ so that a better collection of aggregates is obtained. The rule of thumb we have observed is that if the output has very strong (near 1) coupling measures, one should then try to increase $\delta$ (in order to discover possible subaggregates of those output), and if the output aggregates have weak coupling measures, $\delta$ should be decreased. For this analysis, we will simply use the input $\delta = 0.05$ every time, and report the coupling measures as they first appear.

We generate our random graphs using Algorithm 12 with inputs $n = 1000$ vertices, $m = 10$, 20 or 50 aggregates and $\epsilon = 0.01$, 0.05 or 0.1. We generate 180 graphs in total, 20 with each possible combination of inputs, and then calculate the reversible transition matrix associated with each graph. As the matrices under consideration are random walks on graphs, the stationary distribution of each matrix is known (it is a scalar multiple of the vector of vertex degrees) and so we can make use of the $\pi$-coupling measure in evaluating the outputs.

For each stochastic matrix generated, we calculate the mean and minimum $\pi$-coupling measures of the aggregates output by each application of Algorithm 8 (with input $\delta = 0.05$); as well, we record the number of aggregates output. Then, we take the means of these values over all 20 applications with each pair of inputs. We note that with input $\epsilon$ (into the random graph generator), the $\pi$-coupling measures of the actual aggregates are approximately $1 - \epsilon$. These values are shown in Table C.5.

| | | $\pi$-Coupling | | Mean Number of |
|---|---|---|---|---|
| Aggregates | $1-\epsilon$ | Mean | Min. | Aggregates Output |
| 50 | 0.99 | 0.982 | 0.811 | 50.2 |
| 50 | 0.95 | 0.931 | 0.780 | 47.2 |
| 50 | 0.90 | 0.878 | 0.634 | 29.2 |
| 20 | 0.99 | 0.894 | 0.337 | 22.2 |
| 20 | 0.95 | 0.907 | 0.422 | 20.7 |
| 20 | 0.90 | 0.852 | 0.551 | 12.2 |
| 10 | 0.99 | 0.776 | 0.273 | 12.9 |
| 10 | 0.95 | 0.837 | 0.388 | 11.4 |
| 10 | 0.90 | 0.843 | 0.592 | 8.9 |

Table C.5: The Lower Weighted Algorithm Applied to randomly generated stochastic matrices of order 1000.

We see that with $m = 50$ or 20 aggregates, the algorithm has performed very well; the mean $\pi$-coupling measures of the output aggregates are close to the expected $\pi$-coupling measure of $1-\epsilon$. As well, the minimum $\pi$-coupling measures are significantly lower than these mean values; this suggests that the majority of the aggregates obtained have $\pi$-coupling measures very close $1-\epsilon$. So, a closer examination of the aggregates obtained should allow these outputs to be refined into even stronger $\epsilon$-uncouplings. The outputs obtained from the matrices that have 10 almost invariant aggregates seem to be more problematic. However, in this case we again have the minimum $\pi$-coupling measures significantly lower than the means, suggesting that many of the aggregates obtained are, indeed, almost invariant aggregates.

We chose this particular random matrix generator because it is, in a sense, problematic for our algorithms. Let $A$ be the transition matrix of a random walk on a graph $G$ output by Algorithm 12. For each $i$, let $v(i)$ be the degree of vertex $i$ in the

graph $G$. Then, every entry in the $i$th row of $A$ is equal to either 0 or $1/v(i)$. So, each transition $i \to j$ where $j$ is in a separate aggregate from $i$ is as likely as each transition $i \to j$ where $j$ is in the same aggregate. The Lower Weighted Algorithm (and our other stochastic complement based algorithms) proceeds by assuming that $i$ and $j$ are in the same aggregate whenever

$$a_{ij} = \max_{k \neq i} \{a_{ik}\}.$$

However, this assumption, in this case, is false. We suspect that this is the reason for the low minimum $\pi$-coupling measures, as seen in Table C.5. It seems that some number of states are being assigned to the wrong almost invariant aggregate. However, as we noted above, the mean $\pi$-coupling measure is significantly higher, implying that only a small number of aggregates are being affected in each case. We suspect that the reason for this good performance (on average) is that even though these "cross-aggregate" transitions have the same magnitudes as the transitions within an aggregate, there are far fewer of them. Thus, the likelihood of a correct association being made (in the digraph the algorithm constructs) is quite high. In addition, as the algorithm removes more and more states via stochastic complements, the transitions within aggregates are increased more so than those between aggregates. *i.e.* Incorrect associations are somewhat unlikely, and become even more unlikely as the algorithm removes more states.

# Appendix D

# A low rank example

We examine a particularly simple nearly uncoupled Markov chain, and determine the exact conditions under which the Maximum Entry Algorithm will successfully or unsuccessfully uncouple its associated state space.

Let $v$ and $w$ be entrywise positive column vectors of orders $m$ and $n$, respectively, such that the sum of the entries in each of $v$ and $w$ is 1:

$$v^T \mathbb{1} = w^T \mathbb{1} = 1.$$

Let $p$ and $q$ be positive numbers that are close to, but strictly less than, 1 – for now, we will only assume that $p$ and $q$ are strictly greater than $1/2$. Consider the stochastic matrix

$$A(v, w, p, q) = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}v^T & q\mathbb{1}w^T \end{bmatrix}.$$

The all-ones vectors $\mathbb{1}$ in the first row of blocks have order $m$ and those in the second row have order $n$ – so, the orders of the $(1,1)$ and $(2,2)$th blocks are $m \times m$ and $n \times n$.

The matrix $A(v, w, p, q)$ models the following system. Let $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ where $\mathcal{S}_1$ and $\mathcal{S}_2$ are disjoint, arbitrarily large sets (possibly even infinite). Consider a Markov chain $X$ on $\mathcal{S}$ that satisfies the properties

$$\mathbb{P}\left[x_{t+1} \in \mathcal{S}_1 | x_t \in \mathcal{S}_1\right] = p, \ \mathbb{P}\left[x_{t+1} \in \mathcal{S}_2 | x_t \in \mathcal{S}_1\right] = 1 - p,$$

$$\mathbb{P}\left[x_{t+1} \in \mathcal{S}_2 | x_t \in \mathcal{S}_2\right] = q \text{ and } \mathbb{P}\left[x_{t+1} \in \mathcal{S}_1 | x_t \in \mathcal{S}_1\right] = 1 - q.$$

In addition, for any collections $\mathcal{C}_1 \subseteq \mathcal{S}_1$ or $\mathcal{C}_2 \subseteq \mathcal{S}_2$, the probabilities that $x_t \in \mathcal{C}_1$ or $\mathcal{C}_2$ are dependent only on whether $x_t \in \mathcal{S}_1$ or $\mathcal{S}_2$. For example, given $i \in \mathcal{S}_2$ and $\mathcal{C}_1 \subseteq \mathcal{S}_1$,

$$\begin{aligned}
\mathbb{P}\left[x_{t+1} \in \mathcal{C}_1 | x_t = i\right] &= \mathbb{P}\left[x_{t+1} \in \mathcal{C}_1 | x_t \in \mathcal{S}_2\right] \\
&= \mathbb{P}\left[x_{t+1} \in \mathcal{S}_1 | x_t \in \mathcal{S}_2\right] \mathbb{P}\left[x_{t+1} \in \mathcal{C}_1 | x_{t+1} \in \mathcal{S}_1\right] \\
&= (1-q)\mathbb{P}\left[x_{t+1} \in \mathcal{C}_1 | x_{t+1} \in \mathcal{S}_1\right].
\end{aligned}$$

The vectors $v$ and $w$ are then obtained by discretising or aggregating the collections $\mathcal{S}_1$ and $\mathcal{S}_2$ into disjoint unions

$$\mathcal{S}_1 = \bigcup_{i=1}^{m} \mathcal{E}_i \text{ and } \mathcal{S}_2 = \bigcup_{j=1}^{n} \mathcal{F}_j$$

and defining

$$v_i = \mathbb{P}\left[x_t \in \mathcal{E}_i | x_t \in \mathcal{S}_1\right] \text{ and } w_j = \mathbb{P}\left[x_t \in \mathcal{F}_j | x_t \in \mathcal{S}_2\right].$$

The Markov chain associated with $A = A(v, w, p, q)$ is the Markov chain obtained from $X$ by replacing every member of each of $\mathcal{E}_i$ and $\mathcal{F}_j$ with a single state. For $i = 1, \ldots, m$, state $i$ (of $A(v, w, p, q)$) corresponds to $\mathcal{E}_i$ and for $j = m+1, \ldots, m+n$, state $j$ corresponds to $\mathcal{F}_{j-m}$. Given $v$ and $w$, as above, we define

$$\mathcal{E} = \{1, \ldots, m\} \text{ and } \mathcal{F} = \{m+1, \ldots, m+n\}.$$

Let $A = A(v, w, p, q)$. The collections $\mathcal{E}$ and $\mathcal{F}$ are clearly almost invariant aggregates – the probability of transitioning from one to the other is $1 - p$ or $1 - q$, which we have assumed to be close to 0.

We first define what we mean by a correct output of the Maximum Entry Algorithm. Suppose that $A = A(v, w, p, q)$, $\mathcal{E}$ and $\mathcal{F}$ are as above. Let $G$ be a digraph obtained by applying the Maximum Entry Algorithm to $A$ with input $\delta = 0$. Since $A$ is irreducible, $G$ is weakly connected. Let $i \to j$ be the final directed arc added to $G$ by the Maximum Entry Algorithm and let $G'$ be the subgraph of $G$ obtained by removing the directed arc $i \to j$ (that is, $G'$ is the digraph constructed during the second to last iteration of the algorithm). The subgraph $G'$ has exactly two weakly connected components; we say that $G$ correctly uncouples $A$ if the vertex sets of the weakly connected components of $G'$ are $\mathcal{E}$ and $\mathcal{F}$.

The outputs of the Maximum Entry Algorithm may differ, depending on its implementation – if, at some iteration, the maximum value among the off-diagonal entries

is attained by two or more off-diagonal entries, the exact choice of entry by the algorithm may produce a different output. We say that the Maximum Entry Algorithm correctly uncouples $A$ if every possible implementation produces a digraph $G$ which correctly uncouples $A$.

We have defined the output as correct or incorrect, in this case, for the following reason. Each of our algorithms constructs associations in its output digraph in the same manner. Let $i \to j$ be a directed arc added to the output digraph (by one of our algorithms) and let $\hat{A}$ be the stochastic complement under consideration when this directed arc is added. Then,

$$\hat{a}_{ij} = \max_{k \neq i} \{a_{ik}\}.$$

We are concerned that, under some conditions, the states $i$ and $j$ above may not actually be contained in the same almost invariant aggregate. More importantly, we are concerned that one of our algorithms may remove a state $i$ and add a directed arc $i \to j$ where

1. the state $i$ is contained in an almost invariant aggregate $\mathcal{E}$,

2. $j \notin \mathcal{E}$, and

3. there some state $k \neq i$ (not yet removed) such that $k \in \mathcal{E}$.

If the Maximum Entry Algorithm correctly uncouples the state space of the matrix

$A = A(v, w, p, q)$, then every time the algorithm associates two states $i$ and $j$, either this association is correct or there are no correct associations left to be made.

Investigating this problem provides insight into a fundamental question involved in using our stochastic complement based algorithms – namely, how uncoupled must a matrix be in order for the algorithms to produce correct or reliable output? In Section 5.6.2, we show that if a reversible stochastic matrix is sufficiently near to a block diagonal matrix (in the $\infty$-norm), the Maximum Entry Algorithm will produce a correct output. However, this is an existence-style proof only – it provides no concrete bound or formula. In general, this seems to be a hard problem; however, in this specific case we are able to produce a complete answer.

As well, we are interested in the effect that the discretisation process has on our algorithms. If the discretisation utilised is trivial, namely if $v = w = [1]$, it is impossible to produce an incorrect output. We are interested in seeing if there are discretisation choices that may fool the algorithm, and obfuscate the uncoupled structure of the matrix. As we will see below, as long as there is at least one entry in each of $v$ and $w$ that is sufficiently large (we produce an exact bound), the algorithm will uncouple $A(v, w, p, q)$ correctly.

We are less concerned with the problem of removing an entire almost invariant aggregate. The matrix $A(v, w, p, q)$ is reversible; let $\Pi_v$ and $\Pi_w$ be the diagonal matrices whose $i$th entries are $v_i$ and $w_i$ respectively. Left-multiplication of $A$ by the

matrix

$$\Pi = \begin{bmatrix} (1-q)\Pi_v & 0 \\ 0 & (1-p)\Pi_w \end{bmatrix}$$

obtains a symmetric matrix. As long as every directed arc added to the output digraph has both endpoints contained in the same almost invariant aggregate, the results in Appendix A apply. So, for example, if the order of $v$ is $m$, once the algorithm has removed $m-1$ states from $\mathcal{E} = \{1, \ldots, m\}$, leaving one state $i \in \mathcal{E}$ not yet removed, the stochastic complement $\hat{A}$, at that iteration, will satisfy

$$\hat{a}_{ii} \geq \frac{p^2}{1 + (m-2)(1-p)}.$$

Thus, our Maximum Entry Algorithm (with an appropriate choice of the input $\delta$) can be relied upon to construct stochastic complements that do not remove an entire aggregate.

For the remainder of this section, $A = A(v, w, p, q)$, $\mathcal{E}$ and $\mathcal{F}$ are as defined above. We refer to the submatrices $(1-p)\mathbb{1}w^T$ and $(1-q)\mathbb{1}v^T$ of $A$ as the off-diagonal blocks and the submatrices $p\mathbb{1}v^T$ and $q\mathbb{1}w^T$ as the diagonal blocks.

We start with a technical lemma.

**Lemma D.1.** *Let $A = A(v, w, p, q)$ and let $\mathcal{C}$ be a collection of states properly contained in $\mathcal{E}$. Let $\hat{v}$ be the subvector of $v$ corresponding to $\mathcal{E} \setminus \mathcal{C}$, let $v_{\mathcal{C}}$ be the subvector of $v$ corresponding to $\mathcal{C}$ and let $a = v_{\mathcal{C}}^T \mathbb{1}$. Then,*

$$\left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1} = I + \frac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T$$

*and*

$$A \setminus \mathcal{C} = \begin{bmatrix} \frac{p}{1-pa}\mathbb{1}\hat{v}^T & \frac{1-p}{1-pa}\mathbb{1}w^T \\ \frac{1-q}{1-pa}\mathbb{1}\hat{v}^T & \frac{q-(p+q-1)a}{1-pa}\mathbb{1}w^T \end{bmatrix}.$$

**Proof** We show the first claim by simply multiplying the two matrices together:

$$\begin{aligned} \left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)\left(I + \tfrac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T\right) &= I + \tfrac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T - p\mathbb{1}v_{\mathcal{C}}^T - \tfrac{p^2}{1-pa}\mathbb{1}v_{\mathcal{C}}^T\mathbb{1}v_{\mathcal{C}}^T \\ &= I + \tfrac{p-p(1-pa)-p^2a}{1-pa}\mathbb{1}v_{\mathcal{C}}^T \\ &= I. \end{aligned}$$

(We have $p/(1-pa) > 0$, as both $p$ and $a$ are positive numbers strictly less than 1.)

Next, we express the matrix $A = A(v, w, p, q)$ as

$$A \cong \begin{bmatrix} p\mathbb{1}\hat{v}^T & p\mathbb{1}v_{\mathcal{C}}^T & (1-p)\mathbb{1}w^T \\ p\mathbb{1}\hat{v}^T & p\mathbb{1}v_{\mathcal{C}}^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}\hat{v}^T & (1-q)\mathbb{1}v_{\mathcal{C}}^T & q\mathbb{1}w^T \end{bmatrix}$$

Then,

$$\begin{aligned} A \setminus \mathcal{C} &= \begin{bmatrix} p\mathbb{1}\hat{v}^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}\hat{v}^T & q\mathbb{1}w^T \end{bmatrix} \\ &+ \begin{bmatrix} p\mathbb{1}v_{\mathcal{C}}^T \\ (1-q)\mathbb{1}v_{\mathcal{C}}^T \end{bmatrix} \left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1} \begin{bmatrix} p\mathbb{1}\hat{v}^T & (1-p)\mathbb{1}w^T \end{bmatrix}. \end{aligned}$$

We will calculate the four blocks of the matrix one at a time, making use of the above formula for $(I-p\mathbb{1}v_{\mathcal{C}}^T)^{-1}$ and the fact that $a = v_{\mathcal{C}}^T\mathbb{1}$. The block in the $(1,1)$th position is

$$p\mathbb{1}\hat{v}^T + \left(p\mathbb{1}v_{\mathcal{C}}^T\right)\left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1}\left(p\mathbb{1}\hat{v}^T\right)$$

$$\begin{aligned}
&= p\mathbb{1}\hat{v}^T + p^2\mathbb{1}v_{\mathcal{C}}^T\left(I + \tfrac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T\right)\mathbb{1}\hat{v}^T \\
&= p\mathbb{1}\left(1 + pv_{\mathcal{C}}^T\mathbb{1} + \tfrac{p^2}{1-pa}\left(v_{\mathcal{C}}^T\mathbb{1}\right)^2\right)\hat{v}^T \\
&= p\left(1 + pa + \tfrac{p^2a^2}{1-pa}\right)\mathbb{1}\hat{v}^T \\
&= p\tfrac{(1+pa)(1-pa)+p^2a^2}{1-pa}\mathbb{1}\hat{v}^T \\
&= \tfrac{p}{1-pa}\mathbb{1}\hat{v}^T.
\end{aligned}$$

The block in the $(1,2)$th position is

$$(1-p)\mathbb{1}w^T + \left(p\mathbb{1}v_{\mathcal{C}}^T\right)\left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1}\left((1-p)\mathbb{1}w^T\right)$$

$$\begin{aligned}
&= (1-p)\mathbb{1}w^T + p(1-p)\mathbb{1}v_{\mathcal{C}}^T\left(I + \tfrac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T\right)\mathbb{1}w^T \\
&= (1-p)\mathbb{1}\left(1 + pv_{\mathcal{C}}^T\mathbb{1} + \tfrac{p^2}{1-pa}\left(v_{\mathcal{C}}^T\mathbb{1}\right)^2\right)w^T \\
&= (1-p)\left(1 + pa + \tfrac{p^2a^2}{1-pa}\right)\mathbb{1}w^T \\
&= (1-p)\tfrac{(1+pa)(1-pa)+p^2a^2}{1-pa}\mathbb{1}w^T \\
&= \tfrac{1-p}{1-pa}\mathbb{1}w^T.
\end{aligned}$$

The block in the $(2,1)$th position is

$$(1-q)\mathbb{1}\hat{v}^T + \left((1-q)\mathbb{1}v_{\mathcal{C}}^T\right)\left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1}\left(p\mathbb{1}\hat{v}^T\right)$$

$$
\begin{aligned}
&= (1-q)\mathbb{1}\left(1 + pv_{\mathcal{C}}^T\mathbb{1} + \tfrac{p^2}{1-pa}\left(v_{\mathcal{C}}^T\mathbb{1}\right)^2\right)\hat{v}^T \\
&= \tfrac{1-q}{1-pa}\mathbb{1}\hat{v}^T.
\end{aligned}
$$

(The calculations involved in the $(2,1)$ case are similar to those involved in the $(1,2)$ case – we simply need to replace each $(1-p)$ term with $(1-q)$ and each $w^T$ with $\hat{v}^T$.)

Finally, the entry in the $(2,2)$ position is

$$
\begin{aligned}
& q\mathbb{1}w^T + \left((1-q)\mathbb{1}v_{\mathcal{C}}^T\right)\left(I - p\mathbb{1}v_{\mathcal{C}}^T\right)^{-1}\left((1-p)\mathbb{1}w^T\right) \\
&= q\mathbb{1}w^T + (1-p)(1-q)\mathbb{1}v_{\mathcal{C}}^T\left(I + \tfrac{p}{1-pa}\mathbb{1}v_{\mathcal{C}}^T\right)\mathbb{1}w^T \\
&= \mathbb{1}\left(q + (1-p)(1-q)a + \tfrac{(1-p)(1-q)pa^2}{1-pa}\right)w^T \\
&= \left(q + (1-p)(1-q)a\left(1 + \tfrac{pa}{1-pa}\right)\right)\mathbb{1}w^T \\
&= \left(q + \tfrac{(1-p)(1-q)a}{1-pa}\right)\mathbb{1}w^T \\
&= \tfrac{q(1-pa)+(1-p)(1-q)a}{1-pa}\mathbb{1}w^T \\
&= \tfrac{q-(p+q-1)a}{1-pa}\mathbb{1}w^T.
\end{aligned}
$$

It is somewhat simple to show that if $0 < a < 1$ and $1/2 < p, q < 1$, then $q - (p + q - 1)a > 0$. If we hold $p$ and $a$ constant, the function

$$f = q - (p + q - 1)a = (1 - a)q - (p - 1)a$$

is strictly increasing in $q$. So, since $q > 1/2$,

$$f > \frac{1}{2} - \left(p - \frac{1}{2}\right) a.$$

Then, $p < 1$ implies

$$f > \frac{1}{2} - \frac{1}{2}a = \frac{1}{2}(1 - a).$$

∎

Now, suppose that $A = A(v, w, p, q)$ where at least one of $v$ or $w$ has order 2 or greater. In order for the Maximum Entry Algorithm to proceed correctly, no matter what the implementation, the maximal value among the off-diagonal entries must not occur in an off-diagonal block. That is, for

$$A = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}v^T & q\mathbb{1}w^T \end{bmatrix}$$

and

$$z = \max_{i \neq j} \{a_{ij}\},$$

we must have every entry of $(1-p)w$ and $(1-q)v$ strictly less than $z$. Otherwise, the very first step that the algorithm takes could be incorrect (the very first directed arc that it adds to $G$ may have one endpoint in $\mathcal{E}$ and the other in $\mathcal{F}$). We now show that when this condition is met, the Maximum Entry Algorithm will follow the correct first iteration with a number of further correct iterations.

**Lemma D.2.** *Let $A = A(v, w, p, q)$ where $v$ and $w$ have orders $m$ and $n$, respectively. Let the digraph $G$ be formed by an application of the maximum entry algorithm with input $\delta = 0$ and suppose that the first directed arc $i \to j$ added to $G$ by the algorithm has*

$$i, j \in \mathcal{E} = \{1, \ldots, m\}.$$

*If the maximal off-diagonal value of $A$ is strictly greater than every entry in the off-diagonal blocks of $A$, then the first $m-1$ directed arcs $i \to j$ added to $G$ have $i, j \in \mathcal{E}$.*

**Proof** If the maximal off-diagonal value of $A$ is strictly greater than every entry in the off-diagonal blocks, then the first directed arc $i \to j$ added to $G$ must have either

$$i, j \in \mathcal{E} = \{1, \ldots, m\} \text{ or } i, j \in \mathcal{F} = \{m+1, \ldots, m+n\}.$$

Without loss of generality, we assume that the first directed arc added to $G$ has both endpoints in $\mathcal{E}$. This implies that $m \geq 2$. If $m = 2$, then there is nothing to prove – we have $m - 1 = 1$ and the first directed arc added to $G$ has both entries contained in $\mathcal{E}$. So, suppose that $m \geq 3$.

Let

$$\alpha = \max_{1 \le i \le n} \{v_i\} \text{ and } \beta = \max_{1 \le i \le n} \{w_i\}.$$

For $i, j \in \mathcal{S} = \mathcal{E} \cup \mathcal{F}$, the $ij$th entry of

$$A = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}v^T & q\mathbb{1}w^T \end{bmatrix}$$

is

$$a_{ij} = \begin{cases} pv_j & \text{if } i, j \in \mathcal{E} \\ (1-p)w_{j-m} & \text{if } i \in \mathcal{E} \text{ and } j \in \mathcal{F} \\ qw_{j-m} & \text{if } i, j \in \mathcal{F} \\ (1-q)v_j & \text{if } i \in \mathcal{F} \text{ and } j \in \mathcal{E}. \end{cases}$$

The largest values in the diagonal blocks are $p\alpha$ and $q\beta$, respectively. The largest entries in the off-diagonal blocks are $(1-p)\beta$ and $(1-q)\alpha$. The assumption that the first directed arc added by the Maximum Entry Algorithm has endpoints in $\mathcal{E}$ implies that $m \ge 2$ and that the largest off-diagonal value of $A$ is $p\alpha$. So, we have

$$p\alpha > (1-q)\beta,$$

and either $n = 1$ or $p\alpha \ge q\beta$.

We first prove the following claim: Let $\mathcal{C} \subseteq \mathcal{E}$ be such that $1 \le |\mathcal{C}| \le m-2$ and there is at least one $j \in \mathcal{E} \setminus \mathcal{C}$ with $v_j = \alpha$. Then, the largest off-diagonal entry $\hat{a}_{ij}$ of $A \setminus \mathcal{C}$ has $i, j \in \mathcal{E} \setminus \mathcal{C}$ and $v_j = \alpha$.

Let $\hat{v}$ and $v_{\mathcal{C}}$ be the subvectors of $v$ corresponding to $\mathcal{E} \setminus \mathcal{C}$ and $\mathcal{C}$, respectively, and let $a = v_{\mathcal{C}}^T \mathbb{1}$. Consider

$$\hat{A} = A \setminus \mathcal{C} = \begin{bmatrix} \frac{p}{1-pa}\mathbb{1}\hat{v}^T & \frac{1-p}{1-pa}\mathbb{1}w^T \\ \frac{1-q}{1-pa}\mathbb{1}\hat{v}^T & \frac{q-(p+q-1)a}{1-pa}\mathbb{1}w^T \end{bmatrix}$$

(via Lemma D.1). The vector $\hat{v}$ has order 2 or greater and there is at least one $j$ with $\hat{v}_j = \alpha$. So, the largest off-diagonal entry in the first diagonal block is

$$\frac{p}{1-pa}\alpha$$

and the largest entries in the off-diagonal blocks are

$$\frac{1-p}{1-pa}\beta \text{ and } \frac{1-q}{1-pa}\alpha.$$

As noted above, $p\alpha > (1-p)\beta$; as well, $p, q > 1/2$ implies that $p\alpha > (1-q)\alpha$. Thus, the largest off-diagonal entry in the first diagonal block is strictly greater than every entry of the off-diagonal blocks. So, if the vector $w$ has order equal to 1, the largest off-diagonal value appears only within the first diagonal block. Suppose that, instead, $w$ has order greater than or equal to 2. Above, we noted that $p\alpha \geq q\beta$. The largest entry in the second diagonal block of $\hat{A}$ is

$$\frac{q-(p+q-1)a}{1-pa}\beta < \frac{q}{1-pa}\beta \leq \frac{p}{1-pa}\alpha.$$

Thus, in either case, the largest off-diagonal value of $\hat{A}$ appears only in the first diagonal block. Moreover, any pair $i, j \in \{1, \ldots, m\}$, not yet removed, with

$$\hat{a}_{ij} = \max_{k \neq l} \{\hat{a}_{kl}\}$$

has $v_j = \alpha$.

We now proceed to show by induction on $s$, where $1 \leq s \leq m-1$, that

1. the first $s$ directed arcs added to $G$ have both endpoints contained in $\mathcal{E}$,

2. there is $j \in \mathcal{E}$ with $v_j = \alpha$ not removed during the first $s$ iterations.

Let $i_s \to j_s$ be the directed arc added to $G$ during the $s$th iteration and let

$$\mathcal{C}_s = \{i_1, \ldots, i_s\}.$$

By assumption, we have $i_1 \in \mathcal{E}$ and $v_{j_1} = \alpha$. The state $j_1$ is not removed during the first iteration, so both of the above statements hold. Suppose that $1 \leq s \leq m-2$ and that the two statements hold true for $s$ and let $\hat{A}$ be the stochastic complement formed after the first $s$ iterations. Then, $\mathcal{C}_s$ satisfies our claim above, implying that the largest off-diagonal entry $\hat{a}_{i_{s+1}j_{s+1}}$ of $\hat{A}$ has $i_{s+1}, j_{s+1} \in \mathcal{E} \setminus \mathcal{C}_s$ and $v_{j_{s+1}} = \alpha$. So, the first $s+1$ directed arcs added to $G$ have both endpoints in $\mathcal{E}$ and, after $s+1$ iterations, we have $j_{s+1} \in \mathcal{E} \setminus \mathcal{C}_{s+1}$ and $v_{j_{s+1}} = \alpha$. ∎

We now characterise a sufficient condition under which the Maximum Entry Algorithm will correctly uncouple a matrix $A = A(v, w, p, q)$.

**Proposition D.3.** *Let $v$ and $w$ be entrywise positive vectors each of whose entries' sum is 1, let $\alpha$ and $\beta$ be the maximum entries in $v$ and $w$, respectively, let $p, q \in (1/2, 1)$ and let $A = A(v, w, p, q)$. If*

$$\alpha > \frac{1-p}{p} \text{ and } \beta > \frac{1-q}{q},$$

*then the Maximum Entry Algorithm with input $\delta = 0$ will correctly decouple the state space of $A$.*

**Proof** We note that the fact that $p, q > 1/2$ is actually implied by the other assumptions and need not be assumed. Since $v$ is entrywise positive and $v^T \mathbb{1} = 1$, we have $0 < \alpha \leq 1$. Thus,

$$0 \leq \frac{1-p}{p} < \alpha \leq 1$$

implies that $1/2 < p \leq 1$. The same reasoning applies to $q$.

If both $v$ and $w$ have order equal to 1 (that is, if $v = w = [1]$), there is nothing to prove.

First suppose that exactly one of the vectors has order 1 – without loss of generality, we assume that $w$ has order 1 and that $v$ has order $m \geq 2$. So, $w = [1]$ implies that $\beta = 1$. The largest off-diagonal entry of

$$A = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}v^T & q\mathbb{1}w^T \end{bmatrix} = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1} \\ (1-q)v^T & q \end{bmatrix}$$

is one of the values $p\alpha$, $1-p$ or $(1-q)\alpha$. Since $p, q > 1/2$, $p > 1-q$, implying that $p\alpha > (1-q)\alpha$. As well, the assumption that

$$\alpha > \frac{1-p}{p}$$

implies that $p\alpha > 1-p$. The largest off-diagonal entry in $A$ is $p\alpha$, and this value does not occur in an off-diagonal block. Via Lemma D.2, the first $m-1$ directed arcs added to $G$ will have both endpoints contained in $\mathcal{E} = \{1, \ldots, m\}$. So, after $m-1$ iterations, the weakly connected components of $G$ are $\mathcal{E}$ and $\mathcal{F} = \{m+1\}$. So, in this case, the algorithm correctly decouples $A$ (the Maximum Entry Algorithm, with input $\delta = 0$, adds exactly $m$ directed arcs to $G$).

Now, suppose that the orders of $v$ and $w$ are $m \geq 2$ and $n \geq 2$, respectively. The largest off-diagonal value in the diagonal blocks of

$$A = \begin{bmatrix} p\mathbb{1}v^T & (1-p)\mathbb{1}w^T \\ (1-q)\mathbb{1}v^T & q\mathbb{1}w^T \end{bmatrix}$$

are $p\alpha$ and $q\beta$; the largest entries in the off-diagonal blocks are $(1-p)\beta$ and $(1-q)\alpha$. Without loss of generality, assume that $p\alpha \geq q\beta$. Since $p > 1-q$, $p\alpha > (1-q)\alpha$; since $q > 1-p$, $p\alpha \geq q\beta > (1-p)\beta$. So, the maximal value among the off-diagonal entries of $A$ does not occur in the off-diagonal blocks.

We assume that the maximal entry identified during the first iteration of the Maximum Entry Algorithm is contained in the first diagonal block. By Lemma D.2,

the first $m-1$ directed arcs added to $G$ have both endpoints in $\mathcal{E} = \{1, \ldots, m\}$. By Lemma D.1, the stochastic complement formed after these $m-1$ iterations is

$$\hat{A} = A \setminus \mathcal{C} = \begin{bmatrix} \frac{p}{1-p(1-\alpha)}\alpha & \frac{1-p}{1-p(1-\alpha)}w^T \\ \frac{1-q}{1-p(1-\alpha)}\mathbb{1}\alpha & \frac{q-(p+q-1)(1-\alpha)}{1-p(1-\alpha)}\mathbb{1}w^T \end{bmatrix}.$$

(Every time the Maximum Entry Algorithm removes a state $i \in \mathcal{E}$, there is $j \in \mathcal{E}$, not yet removed, with $v_i \leq v_j$; thus, the final, not yet removed state $j \in \mathcal{E}$ has $v_j = \alpha$. So, the vectors $\hat{v}$ and $v_{\mathcal{C}}$ in the statement of Lemma D.2 satisfy $\hat{v} = [\alpha]$ and $v_{\mathcal{C}}^T \mathbb{1} = 1-\alpha$.)

Now, suppose that the largest off-diagonal value of $\hat{A}$ appears in the second diagonal block and does not appear in the off-diagonal blocks. We then apply Lemma D.2 to $\hat{A}$ to show that the next $n-1$ iterations of the algorithm add directed arcs with both endpoints contained in $\mathcal{F} = \{m+1, \ldots, m+n\}$. Thus, we simply need show that the largest off-diagonal value of $\hat{A}$ does not occur in its off-diagonal blocks.

We note that the above formulae for the entries of $\hat{A}$ have a common denominator of $1 - p(1-\alpha)$. This number is positive, as $p < 1$ and $1 - \alpha < 1$. Thus, we can ignore this denominator and simply find the largest numerator among the off-diagonal entries of $\hat{A}$.

We first show that

$$(1-q)\alpha < (q - (p+q-1)(1-\alpha))\,\beta,$$

thus showing that the largest off-diagonal entry in the second diagonal block is strictly larger than any entry in the $(2,1)$th off-diagonal block. Since $\alpha < 1$,

$$\begin{aligned}
\tfrac{q-(p+q-1)(1-\alpha)}{\alpha} &= \tfrac{1-p+(p+q-1)\alpha}{\alpha} \\
&= \tfrac{1-p}{\alpha} + (p+q-1) \\
&> (1-p) + (p+q-1) \\
&= q.
\end{aligned}$$

As well,

$$\beta > \frac{1-q}{q}$$

implies that $q\beta > 1-q$. So,

$$\begin{aligned}
(q-(p+q-1)(1-\alpha))\,\beta &= \tfrac{q-(p+q-1)(1-\alpha)}{\alpha}\alpha\beta \\
&> q\alpha\beta \\
&> (1-q)\alpha.
\end{aligned}$$

Thus, it remains to show that

$$(1-p)\beta < (q-(p+q-1)(1-\alpha))\,\beta.$$

We note that since $p, q > 1/2$, $p+q-1 > 0$; so,

$$q-(p+q-1)(1-\alpha) = 1-p+\alpha(p+q-1) > 1-p.$$

∎

Recall that we have defined an entry $a_{ij} > 0$ of a nearly uncoupled stochastic matrix to be an error term if $i$ is a member of an almost invariant aggregate and $j$ is not a member of that same almost invariant aggregate.

Let $A = A(v, w, p, q)$, $\alpha = \max\{v_i\}$ and $\beta = \max\{w_j\}$. We have

$$p\alpha = \max_{i,j\in\mathcal{E}}\{a_{ij}\}, \text{ and } q\beta = \max_{i,j\in\mathcal{F}}\{a_{ij}\}.$$

As well, the probabilities $1-p$ and $1-q$ are the probabilities of transitioning from one aggregate to the other. We note that the conditions

$$\alpha > \frac{1-p}{p} \text{ and } \beta > \frac{1-q}{q}$$

are equivalent to the conditions $p\alpha > 1-p$ and $q\beta > 1-q$. Thus, these two conditions are met if and only if every row of the matrix $A$ contains a non-error term that is strictly greater than the sum of the error terms in that row.

The following corollary is direct consequence of Proposition D.3, together with the fact that if $v$ and $w$ are positive vectors of orders $m$ and $n$ whose entries' sum is 1, then

$$\alpha = \max\{v_i\} \geq \frac{1}{m} \text{ and } \beta = \max\{w_j\} \geq \frac{1}{n}.$$

**Corollary D.4.** *Let $v$ and $w$ be entrywise positive vectors, each of whose entries' sum is 1, and let $p, q \in (1/2, 1)$. Let $m$ and $n$ be the orders of $v$ and $w$. If*

$$m < \frac{p}{1-p} \text{ and } n < \frac{q}{1-q},$$

*then the Maximum Entry Algorithm will correctly uncouple the state space of $A = A(v, w, p, q)$.*

As $p$ and $q$ converge to 1 (from below), the terms $p/(1-p)$ and $q/(1-q)$ become arbitrarily large. Thus, very well-decoupled systems can contain large numbers of states, whereas less well-decoupled systems require small numbers of states for a guarantee of success.

# Appendix E

# Complexity

We examine the complexities of some of our algorithms. In particular, we show that every one of our decoupling algorithms has complexity $O(n^3)$, where $n$ is the order of the input matrix.

The complexity of an algorithm is an approximation of the number of floating point operations (flops) required to execute it, typically expressed as a function of the size of the input. A floating point operation is any single binary mathematical operation, or a Boolean comparison $x < y$.

As is typical, we will express the complexity of our algorithms as $O(f(n))$, where $n$ is the order of the input matrix. The meaning of this notation is the following. If Algorithm A has complexity $O(f(n))$, then there is a positive constant $a$ and a positive integer $n'$, such that if the matrix $M$ has order $n \geq n'$, then Algorithm A, applied to $M$, requires at most $af(n)$ floating point operations.

**Lemma E.1.** *Let $A$ be a stochastic matrix of order $n$. Constructing a stochastic complement from $A$ by removing 1 state at a time has complexity $O(n^3)$.*

**Proof** First, we consider removing a single state from a matrix $\hat{A}$ of order $m$. We express

$$\hat{A} \cong \begin{bmatrix} B & v \\ w^T & \hat{a}_{ii} \end{bmatrix}.$$

To form the stochastic complement, we need to construct the matrix

$$\hat{A} \setminus i = B + \frac{1}{1 - \hat{a}_{ii}} v w^T.$$

Rather than use the value $1 - \hat{a}_{ii}$, we will calculate the sum

$$\alpha = \sum_{j \neq i} \hat{a}_{ij} = w^T \mathbb{1} = 1 - \hat{a}_{ii}.$$

(As it avoids subtraction, this calculation avoids a certain type of floating-point error.) This requires at most $m - 1$ flops. This calculation, and the ones below can require fewer than this upper bound if the matrices involved contain entries equal to 0. We then calculate

$$\hat{w} = \frac{1}{\alpha} w = \frac{1}{w^T \mathbb{1}} w,$$

which requires another $m - 1$ or fewer flops. Next, the vector product

$$C = v \hat{w}^T$$

requires $(m-1)^2$ or fewer flops. Finally the sum

$$B + C = B + \frac{1}{1 - \hat{a}_{ii}} v w^T$$

requires another $(m-1)^2$ flops. So, in total, the calculation of the stochastic complement $\hat{A} \setminus i$ requires $2(m-1) + 2(m-1)^2$ or fewer floating point operations. At this point we note that if we had used the term $1 - \hat{a}_{ii}$ rather than calculating the sum above, this would have produced a savings of at most $m-2$ flops. This is insignificant (it has a lower polynomial order than the entire task) and so the extra calculation required by using the sum does not negatively effect performance.

In addition to calculating the stochastic complement itself, we may need to "keep track" of the correspondence between the indices of the newly formed stochastic complement and the original matrix. We store the indices of the matrix $\hat{A}$ in vector form:

$$g = \begin{bmatrix} g_1 & g_2 & \cdots & g_m \end{bmatrix}.$$

The indices of $\hat{A} \setminus i$ are then

$$g \setminus i = \begin{bmatrix} g_1 & \cdots & g_{i-1} & g_{i+1} & \cdots & g_m \end{bmatrix}.$$

The calculation of $g \setminus i$ from $g$ requires $m-1$ memory reassignments. We will assume that a memory reassignment has complexity approximately equal to a single floating point operation.

Thus, the calculation of a stochastic complement of a matrix $A$ of order $n$ which removes $k$ states requires at most

$$\sum_{m=n}^{n-k+1} 2(m-1)^2 + 3m - 2$$

flops. We may disregard the $3m-2$ term, as this number will be insignificant compared to the contribution of the $2(m-1)^2$ term. We calculate

$$\begin{aligned}
\sum_{m=n}^{n-k+1} 2(m-1)^2 &\leq \sum_{m=2}^{n} 2(m-1)^2 \\
&= \sum_{j=1}^{n-1} 2j^2 \\
&= 2\frac{n(n-1)(2n-1)}{6} \\
&< \frac{2n^3}{3}.
\end{aligned}$$

∎

**Lemma E.2.** *Let $A$ be a reversible stochastic matrix of order $n$. Then, the Reorder Algorithm (Algorithm 6) has complexity $O(n^2)$.*

**Proof** The Reorder Algorithm constructs a permutation $f$ such that $A(f, f)$ is lower-weighted. (A matrix $\tilde{A}$ is lower-weighted if $\tilde{a}_{ij} \leq \tilde{a}_{ji}$ whenever $i < j$.) The Reorder Algorithm initialises its data to $r = 1$, $s = 2$ and

$$f = \begin{bmatrix} 1 & 2 & \cdots & n \end{bmatrix}.$$

At each iteration,

1. if $a_{f(r)f(t)} \leq a_{f(t)f(r)}$ for $t = s, s+1, \ldots, n$, the Reorder Algorithm increases $r$ by 1; and

2. if there is $t \geq s$ such that $a_{f(r)f(t)} > a_{f(t)f(r)}$, the algorithm chooses such a value $t'$, increases $s$ by 1, and then permutes the subvector

$$\begin{bmatrix} f(r) & f(r+1) & \cdots & f(t') \end{bmatrix}.$$

As well, whenever the algorithm increases $r$, if the new value of $r$ satisfies $r = s$, it then increases $s$ by 1. Thus, the algorithm maintains the condition $r < s$. The algorithm terminates when it achieves $s = n + 1$.

Checking whether or not $a_{f(r)f(t)} \leq a_{f(t)f(r)}$ for $t = s, s+1, \ldots, n$ and, if this does not hold, selecting an index $t'$ that violates this condition is accomplished simultaneously and requires $3(n - s + 1) < 3n$ or fewer floating point operations. (We start with $t = s$. Calculating $a_{f(r)f(t)} - a_{f(t)f(r)}$ is one flop, checking whether this value is negative is another, iterating $t$ if not is a third flop.)

Permuting the subvector $f(r), \ldots, f(t')$ requires $t' - r + 1 < n$ memory reassignments. We assume that a memory reassignment takes as much calculation power as a flop.

So, increasing $r$ by one requires fewer than $3n$ floating point operations and increasing $s$ by one requires few than $4n$ floating point operations.

Therefore, the number of floating point operations required by the Reorder Algorithm with an input size of $n$ is bounded above by $7n^2$. ∎

**Proposition E.3.** *The complexities of the Maximum Entry, Modified Maximum Entry, Minimum Column, Lower-weighted and Perron-ordered Algorithms are* $O(n^3)$.

**Proof** Let $A$ be a stochastic matrix of order $n$.

Finding the largest off-diagonal entry of stochastic matrix of order $m$, testing whether or not this value exceeds the input and then adding a directed arc to a digraph requires $3m^2$ or fewer floating point operations. These tasks are executed at every iteration of the algorithm, of which there are at most $n-1$. So, the total number of flops required by Maximum Entry Algorithm to execute these tasks is bounded by a polynomial of degree 3.

In order to implement the Modified Maximum Entry and Minimum Column Algorithms, we need to keep track of the number of vertices contained in each weakly connected component of $G$. The vector $m$ (at initialisation) is the column vector of order $n$ that has every entry equal to 1. Whenever the directed arc $i \to j$ is added to $G$, we replace $m_j$ with $m_j+m_i$. The extra calculations involved in keeping track of this vector are insignificant compared to the other operations, so the Modified Maximum Entry Algorithm has complexity equal to that of the Maximum Entry Algorithm.

The Minimum Column Algorithm is very similar. The only task which has significant complexity (of order $m^2$ where $m$ is the size of the current stochastic complement)

is that of finding the index with the smallest column sum, which has complexity equal to that of finding the largest off-diagonal entry.

The Perron-ordered Algorithm proceeds at each iteration by testing each diagonal entry, starting with the last, until it finds indices $i > j$ such that

1. $\hat{a}_{ii} < (1-\epsilon)^2/(1+(m_i-2)\epsilon)$, and

2. $\hat{a}_{ij} \geq \hat{a}_{ik}$ for all $k \neq i, j$.

The complexity of this task is $3m^2$ (where $m$ is the order of the current stochastic complement). It proceeds by checking the first condition, which requires a small constant number of flops, and then, if this holds, it checks the second which requires $3(m-1)$ flops. It may possibly have to check $m-1$ of the diagonal entries, thus requiring approximately $3(m-1)^2$ flops. The extra work involved in the Perron-ordered Algorithm does not increase the order of the complexity.

The Lower-weighted Algorithm has slightly increased complexity. In addition to all the same calculations required by the Perron-ordered Algorithm, it must execute the Reorder Algorithm at every iteration. The extra complexity is bounded by

$$\sum_{m=n}^{2} \frac{3}{2}m^2 \approx \frac{1}{2}n^3.$$

∎

Each of our proposed algorithms is very efficient. Applying them to a matrix $A$ is approximately equal, in complexity, to applying Gauss-Jordan elimination with

pivoting or to solving a well-conditioned eigenproblem on a matrix (see, for example, [11]). Applying one of our algorithms is no more calculation-intensive than merely executing the first step of the SVD or Perron cluster algorithms. Moreover, the eigenproblems that the Perron cluster approach must solve in its first steps are in general *not* well-conditioned. By avoiding these spectral methods altogether, our stochastic complement based algorithms proceed with a finite sequence of simple and well-defined matrix operations.

# Appendix F

# Challenging examples

## F.1 Stationary weights and stochastic complements

Let $A$ be an irreducible nearly uncoupled stochastic matrix on the state space $\mathcal{S}$, let $\pi$ be the stationary distribution of $A$ and let $\mathcal{E} \subseteq \mathcal{S}$ be an almost invariant aggregate. Recall that

$$w_\pi(\mathcal{E}) = \frac{\sum_{i \in \mathcal{E}} \sum_{j \in \mathcal{E}} \pi_i a_{ij}}{\sum_{i \in \mathcal{E}} \pi_i}$$

is the $\pi$-coupling measure of $\mathcal{E}$. We define $\eta_\pi(\mathcal{E}) = 1 - w_\pi(\mathcal{E})$. The value $\eta_\pi(\mathcal{E})$ is the expected probability of transitioning from a state $i \in \mathcal{E}$ to a state $j \notin \mathcal{E}$. If we assume that $\mathcal{E}$ is an almost invariant aggregate, then we may also assume that the value $\eta_\pi(\mathcal{E})$ is close to 0.

We consider the effects that removing states contained in $\mathcal{E}$ can have on the value $\eta_\pi(\mathcal{E})$.

**Proposition F.1.** *Let $A$ be an irreducible stochastic matrix on the state space $\mathcal{S}$ and let $\pi$ be the stationary distribution of $A$; let $\mathcal{E},\mathcal{C} \subseteq \mathcal{S}$ be such that $\mathcal{E} \nsubseteq \mathcal{C}$. Let $\hat{A} = A \setminus \mathcal{C}$, $\eta = \eta_\pi(\mathcal{E})$, with respect to $A$, and $\hat{\eta} = \eta_\pi(\mathcal{E} \setminus \mathcal{C})$, with respect to $\hat{A}$. Let*

$$\nu = \frac{\sum_{i \in \mathcal{E} \cap \mathcal{C}} \pi_i}{\sum_{i \in \mathcal{E}} \pi_i};$$

*then,*

$$\hat{\eta} \leq \frac{1}{1-\nu}\eta.$$

**Proof** First, we show that if $\mathcal{E} \cap \mathcal{C}$ is empty, the $\hat{\eta} \leq \eta$. We express

$$A \cong \begin{bmatrix} C & E_1 & E_2 \\ F_1 & B_{11} & B_{12} \\ F_2 & B_{21} & B_{22} \end{bmatrix} \text{ and } \pi^T \cong \begin{bmatrix} \pi_1^T & \pi_2^T & \pi_3^T \end{bmatrix},$$

where the first position corresponds to $\mathcal{E}$, the second to $\mathcal{C}$ and the third to the remainder $\mathcal{S}$. We calculate

$$\hat{A} = A \setminus \mathcal{C} \cong \begin{bmatrix} C + E_1(I - B_{11})^{-1}F_1 & * \\ * & * \end{bmatrix}$$

(only the first diagonal block is required in our calculation); so

$$1 - \eta = \frac{\pi_1^T C \mathbb{1}}{\pi_1^T \mathbb{1}}$$

and

$$1 - \hat{\eta} = \frac{\pi_1^T \left(C + E_1(I - B_{11})^{-1} F_1\right) \mathbb{1}}{\pi_1^T \mathbb{1}} \geq \frac{\pi_1^T C \mathbb{1}}{\pi_1^T \mathbb{1}} = 1 - \eta.$$

(By Proposition 4.6, the stationary distribution of $\hat{A}$ is a scalar multiple of the appropriate subvector of $\pi$. If we use the stationary distribution of $\hat{A}$, rather than that of $A$ in the above expression of $1 - \hat{\eta}$, this scalar multiple appears in both the numerator and the denominator, and thus does not affect the value of $\hat{\eta}$.)

Next, we show that if $\mathcal{C} \subseteq \mathcal{E}$, the inequality holds. Express

$$A \cong \begin{bmatrix} C_{11} & C_{12} & E_1 \\ C_{21} & C_{22} & E_2 \\ F_1 & F_2 & B \end{bmatrix} \text{ and } \pi^T \cong \begin{bmatrix} \pi_1^T & \pi_2^T & \pi_3^T \end{bmatrix},$$

where the first position corresponds to $\mathcal{E} \setminus \mathcal{C}$, the second to $\mathcal{C}$ and the third to $\mathcal{S} \setminus \mathcal{E}$. We have

$$\eta = \frac{\pi_1^T E_1 \mathbb{1} + \pi_2^T E_2 \mathbb{1}}{\pi_1^T \mathbb{1} + \pi_2^T \mathbb{1}}.$$

We calculate the first row of blocks of the stochastic complement

$$\hat{A} = A \setminus \mathcal{C} \cong \begin{bmatrix} C_{11} + C_{12}(I - C_{22})^{-1} C_{21} & E_1 + C_{12}(I - C_{22})^{-1} E_2 \\ * & * \end{bmatrix};$$

this expression implies that

$$\hat{\eta} = \frac{\pi_1^T \left(E_1 \mathbb{1} + C_{12}(I - C_{22})^{-1} E_2 \mathbb{1}\right)}{\pi_1^T \mathbb{1}} = \frac{\pi_1^T E_1 \mathbb{1} + \pi_1^T C_{12}(I - C_{22})^{-1} E_2 \mathbb{1}}{\pi_1^T \mathbb{1}}.$$

Since $\pi^T A = \pi^T$, we have

$$\pi_2^T = \pi_1^T C_{12} + \pi_2^T C_{22} + \pi_3^T F_2,$$

further implying that

$$\pi_1^T C_{12} \leq \pi_2^T (I - C_{22}).$$

Therefore,

$$\begin{aligned} \hat{\eta} &\leq \frac{\pi_1^T E_1 \mathbb{1} + \pi_2^T (I - C_{22})(I - C_{22})^{-1} E_2 \mathbb{1}}{\pi_1^T \mathbb{1}} \\ &= \frac{\pi_1^T E_1 \mathbb{1} + \pi_2^T E_2 \mathbb{1}}{\pi_1^T \mathbb{1}} \\ &= \frac{\pi_1^T \mathbb{1} + \pi_2^T \mathbb{1}}{\pi_1^T \mathbb{1}} \eta. \end{aligned}$$

Finally, we calculate

$$\begin{aligned} \frac{\pi_1^T \mathbb{1} + \pi_2^T \mathbb{1}}{\pi_1^T \mathbb{1}} &= \frac{\sum\limits_{i \in \mathcal{E}} \pi_i}{\sum\limits_{i \in \mathcal{E} \setminus \mathcal{C}} \pi_i} \\ &= \frac{\sum\limits_{i \in \mathcal{E}} \pi_i}{\sum\limits_{i \in \mathcal{E}} \pi_i - \sum\limits_{i \in \mathcal{E} \cap \mathcal{C}} \pi_i} \\ &= \frac{1}{1 - \nu}, \end{aligned}$$

and we can see that if $\mathcal{C} \subseteq \mathcal{E}$, then $\hat{\eta} \leq \eta / (1 - \nu)$.

Now, suppose that both $\mathcal{E} \cap \mathcal{C}$ and $\mathcal{C} \setminus \mathcal{E}$ are nonempty. Let $\mathcal{C}_1 = \mathcal{C} \cap \mathcal{E}$ and let $\mathcal{C}_2 = \mathcal{C} \setminus \mathcal{E}$. Let $A_1 = A \setminus \mathcal{C}_1$ and let $\eta_1 = \eta_\pi(\mathcal{E})$, with respect to $A_1$. Then, via our conclusions above and the fact that $\hat{A} = A_1 \setminus \mathcal{C}_2$,

$$\hat{\eta} \leq \eta_1 \leq \frac{1}{1 - \nu} \eta.$$

∎

Let $A$ be a nearly uncoupled stochastic matrix with stationary distribution $\pi$. By Proposition F.1, if we are to form a stochastic complement of $A$, then removing states with smaller stationary weights produces a better bound for the inflation of the $\eta$-value and thus better preserves the nearly uncoupled structure of $A$.

However, we can construct a matrix that will "fool" the Maximum Entry Algorithm into removing states that have the very highest stationary weights.

**Lemma F.2.** *Let $A$ be an irreducible stochastic matrix with order greater than or equal to 2 and stationary distribution $\pi$. Let $i \neq j$ be such that $a_{ij}$ is maximal among the off-diagonal entries of $A$. Then,*

$$\pi_i \leq \sum_{k \neq i} \pi_k,$$

*with equality if and only if*

1. *$a_{ii} = 1 - a_{ij}$, further implying that $a_{ik} = 0$ if $k \neq i$ and $k \neq j$, and*

2. *for all $k \neq i$, $a_{ki} = a_{ij}$.*

**Remark.** The stationary distribution $\pi$ of a Markov chain satisfies $\pi^T \mathbb{1} = 1$. Thus, for any index $i$,

$$\sum_{k \neq i} \pi_k = 1 - \pi_i.$$

If we have, as above,

$$\pi_i \leq \sum_{k \neq i} \pi_k,$$

then $\pi_i \leq 1 - \pi_i$, further implying that $\pi_i \leq 1/2$.

**Proof** Since $\pi^T = \pi^T A$, we have

$$\pi_i = \sum_k \pi_k a_{ki},$$

which implies that

$$\pi_i = \sum_{k \neq i} \pi_k \frac{a_{ki}}{1 - a_{ii}}.$$

(Since $A$ is irreducible, $a_{ii} \neq 1$.) The assumption that $a_{ij}$ is maximal among the off-diagonal entries implies that for all $k \neq i$,

$$\frac{a_{ki}}{1 - a_{ii}} \leq \frac{a_{ij}}{1 - a_{ii}} \leq 1$$

(since $a_{ij} \leq 1 - a_{ii}$). Therefore,

$$\pi_i = \sum_{k \neq i} \pi_k \frac{a_{ki}}{1 - a_{ii}} \leq \sum_{k \neq i} \pi_k.$$

Equality occurs if and only if $a_{ij} = 1 - a_{ii}$ and for all $k \neq i$, $a_{ki} = a_{ij}$. ∎

Let $A$ be a nearly uncoupled stochastic matrix with stationary distribution $\pi$. Via Lemma F.2, the best upper bound on the stationary weight of a state $i$ selected for removal by the Maximum Entry Algorithm is $\pi_i \leq 1/2$. So, the inflation of the

$\eta$-value (discussed in Proposition F.1) induced by removing $i$ is bounded above by 2. This is a very insufficient bound, as we will see below.

By Lemma F.2, a $2 \times 2$ stochastic matrix that satisfies $a_{ij}$ maximal and $\pi_i = \pi_j$ is simply any symmetric, irreducible $2 \times 2$ matrix; that is,

$$A = \begin{bmatrix} 1-a & a \\ a & 1-a \end{bmatrix},$$

where $0 < a \leq 1$.

We will next examine a class of stochastic matrices that are particularly problematic for our Maximum Entry Algorithm.

**Definition F.3.** Let $0 < a \leq 1/2$ and let $n \geq 3$. We define $F_n(a)$ to be the $n \times n$ stochastic matrix

$$F_n(a) = \begin{bmatrix} 1-a & & & & & & a \\ a & 1-2a & & & & & a \\ & a & 1-2a & & & & \vdots \\ & & \ddots & \ddots & & & \vdots \\ & & & \ddots & 1-2a & a & \\ & & & & a & 1-a & \end{bmatrix},$$

where every unspecified entry is 0.

For example,

$$F_4(0.3) = \begin{bmatrix} 0.7 & 0 & 0 & 0.3 \\ 0.3 & 0.4 & 0 & 0.3 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0.3 & 0.7 \end{bmatrix} \text{ and } F_5(0.2) = \begin{bmatrix} 0.8 & 0 & 0 & 0 & 0.2 \\ 0.2 & 0.6 & 0 & 0 & 0.2 \\ 0 & 0.2 & 0.6 & 0 & 0.2 \\ 0 & 0 & 0.2 & 0.6 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix}.$$

**Proposition F.4.** *Let $C$ be an irreducible stochastic matrix of order $n \geq 3$ and let $\pi$ be the unique stationary distribution of $C$. Suppose that the Maximum Entry Algorithm with input $\delta = 0$ has been applied to $C$; for $k = 1, \ldots, n-1$, let $i_{n+1-k}$ be the state removed by the algorithm during its $k$th iteration; let $i_1$ be the state not removed by the algorithm (during any iteration). Then, for $l = 2, \ldots, n$,*

$$\pi_{i_l} \leq \pi_{i_1} + \ldots + \pi_{i_{l-1}}.$$

*Moreover, we have equality for every $l$ if and only if for some positive number $a \leq 1/2$, we have either $C = F_n(a)$, or permuting indices 1 and 2 of $C$ obtains $F_n(a)$.*

**Proof** Let $C$ and $\pi$ be as above. Without loss of generality, we assume that the algorithm removes state $n$ first, state $n-1$ second, and so forth. Let $C^{(0)} = C$ and let $C^{(k)}$ be the stochastic complement formed during the $k$th iteration of the algorithm. Let $\pi^{(k)}$ be the stationary distribution of $C^{(k)}$. By Proposition 4.6,

$$\pi^{(k)} = \alpha \begin{bmatrix} \pi_1 & \cdots & \pi_{n-k} \end{bmatrix},$$

where $\alpha$ is chosen so that $(\pi^{(k)})^T \mathbb{1} = 1$. Because the Maximum Entry Algorithm removes state $n+1-k$ during iteration $k$, the largest off-diagonal value of $C^{(k)}$ is contained in the bottom ($(n-k)$th) row. Via Lemma F.2, for $k = 0, \ldots, n-2$, we have

$$\pi_{n-k}^{(k)} \leq \pi_1^{(k)} + \ldots + \pi_{n-k-1}^{(k)},$$

further implying that if $0 \leq k \leq n-2$,

$$\pi_{n-k} \leq \pi_1 + \ldots + \pi_{n-k-1}.$$

We now show that the statement concerning equality holds for $n = 3$ and then proceed by induction on $n$.

For $n = 3$, the first directed arc added to the output digraph by the algorithm must be $3 \to 2$ or $3 \to 1$ (since we have assume the algorithm removes state 3 first). First, suppose that the directed arc $3 \to 2$ is the first directed arc added and let $a = c_{32}$. Thus, by Lemma F.2, we have

$$C = \begin{bmatrix} c_{11} & c_{12} & a \\ c_{21} & c_{22} & a \\ 0 & a & 1-a \end{bmatrix}.$$

This form alone guarantees that $\pi_3 = \pi_1 + \pi_2$. Moreover, the fact that $3 \to 2$ is added first implies that $c_{12}, c_{21} \leq a$. In order to have $\pi_2 = \pi_1$, the matrix $C \setminus 3$ must be symmetric. We calculate

$$C \setminus 3 = \begin{bmatrix} c_{11} & c_{12} & a \\ c_{21} & c_{22} & a \\ 0 & a & 1-a \end{bmatrix} + \frac{1}{1-(1-a)} \begin{bmatrix} a \\ a \end{bmatrix} \begin{bmatrix} 0 & a \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12}+a \\ c_{21} & c_{22}+a \end{bmatrix}.$$

So, $\pi_3 = \pi_2 + \pi_1$ and $\pi_2 = \pi_1$, if and only if, in addition to the conditions set forth in Lemma F.2, we have $c_{12} + a = c_{21}$. We note that this implies that $c_{21} \geq a$; since $c_{32} = a$ is maximal, we must have $c_{21} = a$, in turn implying that $c_{12} = 0$. Then, we simply solve for the diagonal entries and we see that $C$ must be

$$C = \begin{bmatrix} 1-a & 0 & a \\ a & 1-2a & a \\ 0 & a & 1-a \end{bmatrix} = F_3(a).$$

If we suppose that the first directed arc added by the algorithm is $3 \to 1$, we obtain, in a very similar manner,

$$C = \begin{bmatrix} 1-2a & a & a \\ 0 & 1-a & a \\ a & 0 & 1-a \end{bmatrix} \cong F_3(a).$$

(Permuting indices 1 and 2 of this matrix obtains $F_3(a)$.) We note that we must have $0 < a \leq 1/2$, in either case – if $a = 0$, the matrix is reducible and if $a > 1/2$, one of the diagonal entries is negative.

Now, suppose that $n \geq 4$; suppose further that

$$\pi_{n-k} = \pi_1 + \ldots + \pi_{n-k-1},$$

for $k = 0, \ldots, n-2$, and that the statement of the proposition holds for $n' \leq n-1$.

By assumption (since state $n$ is removed first), the largest value among the off-diagonal entries occurs in the bottom row of $C$, say $c_{nj} = a$ is maximal (where $j \neq n$). By Lemma F.2, $c_{nn} = 1 - a$ and $c_{kn} = a$ for all $k \neq n$. The matrix $C$ has the form

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1,n-1} & a \\ \vdots & \ddots & \vdots & \vdots \\ c_{n-1,1} & \cdots & c_{n-1,n-1} & a \\ c_{n,1} & \cdots & c_{n,n-1} & 1-a \end{bmatrix},$$

where exactly one of the numbers $c_{n1}, \ldots, c_{n,n-1}$ is equal to $a$ and the remainder are equal to 0.

Let $\hat{C} = C \setminus n$. By the inductive hypothesis, either $\hat{C} = F_{n-1}(a')$ or swapping indices 1 and 2 of $\hat{C}$ obtains $F_{n-1}(a')$ for some positive $a' \leq 1/2$. In either case,

$$\frac{c_{kn}c_{n,n-1}}{1 - c_{nn}} = \hat{c}_{k,n-1} = a',$$

for all $k \leq n-2$. This implies that, $c_{n,n-1} \neq 0$, and so we see that for all $k \leq n-2$, $c_{nk} = 0$. Thus,

$$C = \begin{bmatrix} B & a\mathbb{1} \\ ae_{n-1}^T & 1-a \end{bmatrix},$$

where $e_{n-1}$ is the vector of order $n-1$ with $(n-1)$th entry equal to 1 and every other entry equal to 0. Now, permuting the first and second indices of

$$\hat{C} = B + \frac{1}{1-(1-a)}(a\mathbb{1})(ae_{n-1}^T) = C + a\mathbb{1}e_{n-1}^T$$

leaves the matrix $a\mathbb{1}e_{n-1}^T$ fixed, as $n-1 \geq 4$. So, either

$$B = F_{n-1}(a') - a\mathbb{1}e_{n-1}^T,$$

or permuting the first and second indices of $B$ obtains this matrix. We calculate

$$F_{n-1}(a') - a\mathbb{1}e_{n-1}^T = \begin{bmatrix} 1-a' & & & & & & a'-a \\ a' & 1-2a' & & & & & a'-a \\ & a' & 1-2a' & & & & \vdots \\ & & \ddots & \ddots & & & \vdots \\ & & & \ddots & 1-2a' & a'-a & \\ & & & & a' & 1-a'-a & \end{bmatrix}.$$

The matrix $B \cong F_{n-1}(a') - a\mathbb{1}e_{n-1}^T$ is a principal submatrix of $C$. So, we see that $C$ has off-diagonal entries equal to $a'-a$, implying that $a' \geq a$. As well, $C$ has off-diagonal entries equal to $a'$; we have assumed that the value $a$ is maximal among the off-diagonal entries, so we have $a' \leq a$. Therefore, in fact, $a' = a$. Thus, either

$$B = F_{n-1}(a) - a\mathbb{1}e_{n-1}^T = \begin{bmatrix} 1-a & & & & & \\ a & 1-2a & & & & \\ & a & 1-2a & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & 1-2a & \\ & & & & a & 1-2a \end{bmatrix},$$

or permuting the first two indices of $B$ obtains this matrix. When we insert $B$ into the formula for $C$, above, we see that $C \cong F_n(a)$, either identically or by permuting positions 1 and 2. ∎

The stationary distribution $\pi$ of $F_n(a)$ satisfies

$$\begin{aligned} \pi_n &= \pi_1 + \ldots + \pi_{n-1} \\ \pi_{n-1} &= \pi_1 + \ldots + \pi_{n-2} \\ &\vdots \\ \pi_2 &= \pi_1. \end{aligned}$$

A very simple proof by induction shows that the stationary distribution $\pi$ of $F_n(a)$ is

$$\pi^T = \begin{bmatrix} \frac{1}{2^{n-1}} & \frac{1}{2^{n-1}} & \frac{1}{2^{n-2}} & \cdots & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

We note that for $2 \leq l \leq n$, and $\pi$ as above,

$$\sum_{k=l}^{n} \pi_k = \frac{2^{l-2} + \ldots + 2^{n-2}}{2^{n-1}} = \frac{2^{n-1} - 2^{l-1}}{2^{n-1}} = 1 - \frac{1}{2^{n-l}}.$$

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$ and let $\pi$ be the stationary distribution of $A$. Suppose that there is an almost invariant aggregate $\mathcal{E} \subseteq \mathcal{S}$ such that the principal submatrix $C = A(\mathcal{E})$ is a small perturbation of $F_n(a)$, where $n$ is large. That is, we assume that $\|C - F_n(a)\|$ is small, using some appropriate matrix norm.

Since $C$ is approximately equal to $F_n(a)$, we assume that the Maximum Entry Algorithm applied to $C$, removes states contained in $\mathcal{E}$ in approximately the same order as it would if applied to $F_n(a)$. As well, we assume that the subvector $\pi(\mathcal{E})$ is approximately equal to a scalar multiple of the stationary distribution of $F_n(a)$. That is, it may be that

$$\pi_{i_1} \approx \frac{\alpha}{2^{n-1}}$$

and, for $k \geq 2$,

$$\pi_{i_k} \approx \frac{\alpha}{2^{n+1-k}},$$

where $\alpha = \pi(\mathcal{E})^T \mathbb{1}$. Moreover, the Maximum Entry Algorithm may remove state $i_n$ first, state $i_{n-1}$ second, and so forth. Assume that this is in fact the case. After removing $s$ members of $\mathcal{E}$, the best upper bound on the inflation of $\eta_\pi(\mathcal{E})$ (discussed above) is

$$\hat{\eta} \leq \frac{1}{1-\nu}\eta,$$

where

$$\nu = \frac{\sum_{k=n-s}^{n} \pi_{i_k}}{\sum_{k=1}^{n} \pi_{i_k}} \approx 1 - \frac{1}{2^s}$$

(via our above calculation with $l = n - s$ and the fact that $\pi(\mathcal{E})^T \mathbb{1} = \alpha$). So,

$$\hat{\eta} \leq 2^s \eta.$$

Even if the value $\eta$ is insignificant, for sufficiently large $s$, this upper bound can be become quite large. For example, consider the $31 \times 31$ nearly uncoupled stochastic matrix

$$A = \begin{bmatrix}
0.5 & & & & & & * & \epsilon \\
0.5+\delta & 0 & & & & & * & \epsilon \\
 & 0.5+2\delta & 0 & & & & * & \epsilon \\
 & & \ddots & \ddots & & & \vdots & \vdots \\
 & & & \ddots & 0 & & * & \epsilon \\
 & & & & 0.5+29\delta & & * & \epsilon \\
\epsilon & 0 & 0 & \cdots & \cdots & 0 & 1-\epsilon &
\end{bmatrix},$$

where every unspecified entry is equal to 0, $\delta = 10^{-6}$, $\epsilon = 10^{-7}$ and the $*$ entries are chosen so that each row sum is 1. This matrix is very clearly uncoupled, the probability of transitioning from any member of $\mathcal{E} = \{1, \ldots, 30\}$ to state 31, and *vice versa*, is

$$\epsilon = 0.0000001.$$

The principal submatrix on states 1 through 30 is a small perturbation of $F_{30}(0.5)$.

We have chosen this exact perturbation so that any implementation of the Maximum Entry Algorithm removes state 30 first, state 29 second, and so forth, until it removes state 2, leaving states 1 and 31. At this point we would hope that the algorithm terminates – every association made so far has been correct and the remaining states are distinct representatives of the almost invariant aggregates. However, our calculations below show that unless the input value for Algorithm 3 has been chosen very conservatively (approximately 0.5 or greater), the algorithm will proceed to remove state 1, making an error.

Calculation using Matlab shows that

$$A \setminus \{2, \ldots, 30\} \approx \begin{bmatrix} 0.5046 & 0.4954 \\ \epsilon & 1-\epsilon \end{bmatrix}.$$

This does not attain the inflation by $2^{29}$, but it is somewhat close. The probability of transitioning from state 1 to state 31 has been increased by a factor of approximately $4.954 \times 10^6$. Thus, the average inflation of this transition probability by these 29 stochastic complements is

$$\sqrt[29]{4.954 \times 10^6} \approx 1.7016.$$

After 29 stochastic complements, the Maximum Entry Algorithm has failed to preserve the nearly uncoupled structure of the matrix. On average, the $\eta$-value has been inflated by a factor of approximately 1.7 by each successive stochastic complement.

Calculation shows that the Minimum Column Algorithm, applied to the above matrix $C$, removes the states 1 through 29 in the order

$$2,\ 3,\ \ldots, 16,\ 1,\ 17,\ 18,\ \ldots,\ 29.$$

After removing these 29 states, the stochastic complement formed is

$$\tilde{C} = \begin{bmatrix} 1 - z_1 & z_1 \\ z_2 & 1 - z_2 \end{bmatrix},$$

where

$$z_1 \approx 2.0001 \times 10^{-7} \approx 2\epsilon$$

and

$$z_2 \approx \epsilon - 2\epsilon^2.$$

The probability of transitioning from state 30 to state 31 is approximately doubled by these 29 complements and the probability of transitioning from state 31 to state 31 has been fractionally increased; the nearly uncoupled structure of the matrix has been preserved, more or less.

The Minimum Column Algorithm tries to reduce error (prevent large inflation of the $\eta$-value) by removing states with low column sums, rather than low stationary

weights; thus, it did not remove the states in the order determined by the vector $\pi$, above. We suspect that small stationary weight is a better choice than small column sum in attempting to prevent error inflation – however, the Minimum Column algorithm compares very well with other methods and in this case produces a far superior output.

We have presented the Maximum Entry Algorithm as it is intuitive, simple and, in examples based on practical data, performs very well (usually, as well as any other algorithm we have examined). However, we suggest that care needs to be utilised in its application, as exotic structures in the matrix or the associated digraph seem to mislead it. We suspect that the error-reducing algorithms are the most robust.

## F.2 Paths and cycles

We examine two classes of substochastic matrices that are problematic for all of our uncoupling algorithms – namely long paths and cycles.

Let $B$ be a substochastic matrix on the states $\mathcal{C}$; we refer to $B$ as *path-like* if

$$B \cong \begin{bmatrix} * & * & & \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ & & * & * \end{bmatrix},$$

where the unspecified entries are zeroes. A path-like matrix is more commonly referred

to as tridiagonal. We refer to $B$ as *cyclic* if

$$B \cong \begin{bmatrix} * & * & & & * \\ * & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & * \\ * & & & * & * \end{bmatrix},$$

where the unspecified entries are 0. That is, $B$ is cyclic if its indices can be ordered into $\{i_1, \ldots, i_n\}$ such that $b_{i_k i_l} \neq 0$ only if $l = k$ or $l \equiv k \pm 1$ *modulo* $n$.

Let $A$ be a nearly uncoupled stochastic matrix on the state space $\mathcal{S}$; suppose that there is an almost invariant aggregate $\mathcal{E} \subseteq \mathcal{S}$ such that the principal submatrix $B = A(\mathcal{E})$ is cyclic or path-like. We claim that constructing the almost invariant aggregates of such a matrix is very problematic. As well, via continuity, if $B = A(\mathcal{E})$ is a small perturbation of a cyclic or path-like matrix similar problems arise.

The problem that cyclic or path-like examples poses for our stochastic complement based algorithms is, in a sense, the opposite of the problem we encountered with the $F_n(a)$ matrices. Namely, large numbers of stochastic complements on cyclic and path-like examples can drastically shrink significant entries. (The problem with the $F_n(a)$ matrices is that large numbers of poorly chosen complements can greatly increase insignificant entries.)

**Lemma F.5.** *Let $n \geq 1$ be a positive integer and let $0 < a < 1/2$. Let $P_n$ be the*

$n \times n$ *matrix*

$$P_n = \begin{bmatrix} 1 & -a & & \\ -a & 1 & \ddots & \\ & \ddots & \ddots & -a \\ & & -a & 1 \end{bmatrix}$$

*where every unspecified entry is equal to* 0 *(for* $n = 1$, $P_n = [1]$*). Let* $\Delta_n$ *be the determinant of* $P_n$*. Then,*

1. *the values* $\Delta_n$ *are strictly decreasing in* $n$*;*

2. *we have* $\Delta_1 = 1$, $\Delta_2 = 1 - a^2$ *and, for* $n \geq 3$,

$$\Delta_n = \Delta_{n-1} - a^2 \Delta_{n-2};$$

3. *for* $n \geq 2$, *the* $(1,1)$*th and* $(n,n)$*th entries of* $P_n^{-1}$ *are both equal to*

$$\frac{\Delta_{n-1}}{\Delta_n}$$

*and the* $(1,n)$*th and* $(n,1)$*th entries of* $P_n^{-1}$ *are both equal to*

$$\frac{a^{n-1}}{\Delta_n};$$

4. *moreover, the above entries satisfy*

$$\lim_{n\to\infty} \frac{\Delta_{n-1}}{\Delta_n} = \frac{1 - \sqrt{1 - 4a^2}}{2a^2}, \text{ and } \lim_{n\to\infty} \frac{a^{n-1}}{\Delta_n} = 0.$$

**Proof** For any $n$, the matrix $P_n$ is symmetric and so its eigenvalues are real numbers. A simple application of the Perron-Frobenius Theorem shows that the Perron value of $I - P_n$ is less than or equal to $2a$ – thus, every eigenvalue $\lambda$ of $P_n$ satisfies

$$1 - \lambda \leq |1 - \lambda| \leq 2a < 1.$$

This implies that every eigenvalue of $P_n$ is positive; therefore, for all $n$, $\Delta_n$ is positive.

The second statement is obtained from a formula for the determinant of a tridiagonal matrix found in [14, Section 0.9.10]. The first statement is a consequence of the second together with the fact that each $\Delta_n$ is positive.

The third statement can be obtained from the well-known Cramer's rule; for example, see [14, Sections 0.8.3 and 0.8.4].

Now, for $n \geq 2$, we define

$$\rho_n = \frac{\Delta_{n-1}}{\Delta_n}.$$

Since the terms $\Delta_n$ are strictly decreasing in $n$, the terms $\rho_n$ are strictly increasing in $n$.

We note that for $n \geq 3$,

$$\frac{1}{\rho_n} = \frac{\Delta_n}{\Delta_{n-1}} = \frac{\Delta_{n-1} - a^2\Delta_{n-2}}{\Delta_{n-1}} = 1 - a^2\rho_{n-1}.$$

First, we use induction on $n \geq 2$ to show that $\rho_n < 1/a$. For $n = 2$, we note that $0 < a < 1/2$ implies that $a < 1 - a^2$; so,

$$\rho_2 = \frac{1}{1-a^2} < \frac{1}{a}.$$

Now, if $n \geq 3$ and $\rho_{n-1} < 1/a$, then

$$\frac{1}{\rho_n} = 1 - a^2\rho_{n-1} > 1 - a > a$$

(again, via the fact that $0 < a < 1/2$). Therefore, for all $n \geq 2$, $\rho_n < 1/a$.

Since $\rho_n$ is positive, increasing in $n$, and bounded above by $1/a$, we have

$$\lim_{n\to\infty} \rho_n = \rho \leq \frac{1}{a}.$$

for some positive real number $\rho$. If we apply this limit to the equality

$$\frac{1}{\rho_n} = 1 - a^2\rho_{n-1},$$

we see that

$$\frac{1}{\rho} = 1 - a^2\rho,$$

further implying that

$$\rho = \frac{1 \pm \sqrt{1-4a^2}}{2a^2}.$$

We note

$$\begin{aligned}
\tfrac{1+\sqrt{1-4a^2}}{2a^2} &= \tfrac{1+\sqrt{(1-2a)(1+2a)}}{2a^2} \\
&> \tfrac{1+\sqrt{(1-2a)^2}}{2a^2} \\
&= \tfrac{1-a}{a^2} \\
&\geq \tfrac{a}{a^2} \\
&= \tfrac{1}{a}.
\end{aligned}$$

As we noted above, $\rho \leq 1/a$, so we must have

$$\rho = \frac{1-\sqrt{1-4a^2}}{2a^2}$$

In fact, $\rho < 1/a$:

$$\begin{aligned}
\tfrac{1-\sqrt{1-4a^2}}{2a^2} &= \tfrac{1-\sqrt{(1-2a)(1+2a)}}{2a^2} \\
&< \tfrac{1-\sqrt{(1-2a)^2}}{2a^2} \\
&= \tfrac{1}{a}.
\end{aligned}$$

Finally, for $n \geq 2$,

$$\frac{a^{n-1}}{\Delta_n} = \frac{\Delta_{n-1}}{\Delta_n}\frac{a^{n-1}}{\Delta_{n-1}} = \rho_n \frac{a^{n-1}}{\Delta_{n-1}}.$$

Therefore, for all $n \geq 1$,

$$\frac{a^{n-1}}{\Delta_n} = \frac{\rho_2 \ldots \rho_n a^{n-1}}{\Delta_1} < \frac{(a\rho)^{n-1}}{\Delta_1}$$

(since $\rho_n$ is strictly increasing, $\rho_n < \rho$ for all $n$). As we noted above, $\rho < 1/a$ and so $a\rho < 1$; the sequence $a^{n-1}/\Delta_n$ converges to 0. ∎

Now, let $A$ be a stochastic matrix that is nearly uncoupled with respect to $\epsilon > 0$. Suppose that there is a principal submatrix $B$ of $A$ of the form

$$B = \begin{bmatrix} \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & & & \\ \frac{1-\epsilon}{2} & 0 & \frac{1-\epsilon}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1-\epsilon}{2} & 0 & \frac{1-\epsilon}{2} \\ & & & \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} \end{bmatrix}.$$

Without loss of generality we assume that the states associated with the above expression of $B$ are $\{1, \ldots, n\}$; we further suppose that $n$ is very large. The states associated with $B$ form a minimal almost invariant aggregate. Consider the effects of removing states 2 through $n-1$ via stochastic complements.

$$\hat{B} = B \setminus \{j : 2 \leq j \leq n-1\} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$$

$$+\begin{bmatrix} a & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & a \end{bmatrix} \begin{bmatrix} 1 & -a & & & \\ -a & 1 & \ddots & & \\ & \ddots & \ddots & -a & \\ & & -a & 1 & \end{bmatrix}^{-1} \begin{bmatrix} a & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & a \end{bmatrix},$$

where the order of the square matrix in the second term is $n-2$ and

$$a = \frac{1-\epsilon}{2}.$$

Via Lemma F.5,

$$\hat{B} = \begin{bmatrix} a + a^2 \frac{\Delta_{n-3}}{\Delta_{n-2}} & a^2 \frac{a^{n-3}}{\Delta_{n-2}} \\ a^2 \frac{a^{n-3}}{\Delta_{n-2}} & a + a^2 \frac{\Delta_{n-3}}{\Delta_{n-2}} \end{bmatrix},$$

where $\Delta_k$ is as defined in Lemma F.5. We note that, again via the above lemma, for $n$ sufficiently large, the off-diagonal terms of $\hat{B}$ vanish and the diagonal terms are approximately equal to

$$a + a^2 \frac{1-\sqrt{1-4a^2}}{2a^2} = \frac{1-\epsilon}{2} + \frac{1-\sqrt{2\epsilon-\epsilon^2}}{2}.$$

When $\epsilon$ is small, the above expression is well approximated by $1-\sqrt{\epsilon/2}$. So, our above assumptions imply that

$$\hat{B} \approx \begin{bmatrix} 1-\sqrt{\epsilon/2} & 0 \\ 0 & 1-\sqrt{\epsilon/2} \end{bmatrix}.$$

The above sequence of complements has not altered the fact that $\hat{B}$ corresponds to an almost invariant aggregate, but this aggregate is no longer minimal (with respect to the Markov chain induced by the stochastic complement).

For example, let $n = 60$, $\epsilon = 0.01$ and consider the substochastic matrix $B$, as above. We remove all of the interior members of the path through stochastic complements (the calculation is accomplished with MatLab):

$$B \setminus \{2, \ldots, 59\} \approx \begin{bmatrix} 0.8189 & 1.649 \times 10^{-9} \\ 1.649 \times 10^{-9} & 0.8189 \end{bmatrix}$$

The off-diagonal terms (which represent transitions within an aggregate) are significantly smaller than $\epsilon$; the diagonal terms are somewhat close to $1 - \sqrt{\epsilon/2} \approx 0.8419$.

The cause of this behaviour is intuitively simple to understand. Any two states within a path-like aggregate are connected by a sequence of significant transitions. However, if the Markov chain starts near one of the ends of the path, the expected number of transitions before visiting the other end can be quite large. The Markov chain tends to wander back and forth along sections of the path, and can take a great deal of time before it visits two states that are very far apart. Such separated states are not, in fact, well-connected to each other, and removing large numbers of states from in-between, via stochastic complements, simply makes this poor connection apparent.

In general, if one of our algorithms removes a large numbers of consecutive states from a path-like aggregate, it can become unlikely that the algorithm will correctly

associate the remainder of the path.

Cyclic aggregates are problematic in exactly the same manner. A cycle is a cyclic sequence of paths; removing large numbers of consecutive states results in the exact same terms we saw above. For example, suppose that $n$ is very large and let $B$ be the $3n \times 3n$ cyclic substochastic matrix with

$$b_{i,i+1} = b_{i+1,i} = b_{3n,1} = b_{1,3n} = \frac{1-\epsilon}{2} = a$$

$(1 \leq i \leq 3n-1)$ and all other terms equal to 0. Removing states

$$\mathcal{C} = \{2, \ldots, n, n+2, \ldots, 2n, 2n+2, \ldots, 3n\}$$

(leaving states 1, $n+1$ and $2n+1$) constructs the stochastic complement

$$\hat{B} = B \setminus \mathcal{C} = \begin{bmatrix} 2a^2\frac{\Delta_{n-2}}{\Delta_{n-1}} & a^2\frac{a^{n-2}}{\Delta_{n-1}} & a^2\frac{a^{n-2}}{\Delta_{n-1}} \\ a^2\frac{a^{n-2}}{\Delta_{n-1}} & 2a^2\frac{\Delta_{n-2}}{\Delta_{n-1}} & a^2\frac{a^{n-2}}{\Delta_{n-1}} \\ a^2\frac{a^{n-2}}{\Delta_{n-1}} & a^2\frac{a^{n-2}}{\Delta_{n-1}} & 2a^2\frac{\Delta_{n-2}}{\Delta_{n-1}} \end{bmatrix}.$$

As $n \to \infty$, the off-diagonal terms of such a matrix approach 0 and the diagonal terms approach

$$2a^2\frac{1-\sqrt{1-4a^2}}{2a^2} = 1 - \sqrt{2\epsilon - \epsilon^2} \approx 1 - \sqrt{2\epsilon}.$$

Suppose that $B$ is a principle submatrix of some larger stochastic matrix $A$. Evidently, for $n$ sufficiently large, removing the collection of states $\mathcal{C}$ may split the minimal

almost invariant aggregate into 3. For example, we produce the matrix $\hat{B}$ for $n = 30$ and $\epsilon = 0.01$:

$$\hat{C} = \begin{bmatrix} 0.8589 & 0.0020 & 0.0020 \\ 0.0020 & 0.8589 & 0.0020 \\ 0.0020 & 0.0020 & 0.8589 \end{bmatrix}$$

The off-diagonal terms are again less than $\epsilon$ and the diagonal terms are very close to $1 - \sqrt{2\epsilon} \approx 0.8586$.

If we apply one of our stochastic complement based algorithms to a nearly uncoupled stochastic that has a cyclic almost invariant aggregate, this aggregate is likely to be "split" into subaggregates, and not correctly linked in the output digraph.

The Perron cluster approach is problematic with regards to long cycles and paths, in an interestingly similar manner.

Let $Q_n$ be the adjacency matrix of the undirected cycle on $n \geq 2$ vertices; that is, $Q_n$ is the $(0, 1)$-matrix

$$Q_n = \begin{bmatrix} 0 & 1 & & 1 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & 0 \end{bmatrix}$$

of order $n$. The eigenvalues of $Q_n$ are

$$\lambda_k = 2\cos\left(\frac{2(k-1)\pi}{n}\right),$$

for $k = 1, 2, \ldots, n$ [4, Section 1.2]. When $n$ is particularly large the matrix $Q_n$ possesses multiple eigenvalues that are near 2:

$$2\cos(0) = 2 \text{ and } 2\cos(\frac{2\pi}{n}) \approx 2\cos(0) = 2$$

(and others). Thus, when $n$ is large, the irreducible, reversible substochastic matrix

$$B = \frac{1-\epsilon}{2} Q_n$$

possesses multiple eigenvalues that are very near to $1 - \epsilon$. So, if the stochastic matrix $A$ is nearly uncoupled and possesses $B$ as a principal submatrix, the submatrix $B$ may contribute multiple eigenvalues to the Perron cluster. An assumption in the reasoning behind both the PCCA and PCCA+ Algorithms is that each almost invariant aggregate has a principal submatrix with exactly one eigenvalue near to 1.

The problem that these specific eigenvalues pose to these approaches is very similar to what we saw above; namely, these algorithms can split such an aggregate into smaller subaggregates. For example, an eigenvector of $Q_n$ associated with the eigenvalue $2\cos(2\pi/n)$ is

$$v = \begin{bmatrix} \cos(2\pi/n) & \cos(4\pi/n) & \cos(8\pi/n) & \cdots & \cos(2n\pi/n) \end{bmatrix}.$$

We note that if $k \approx n/2$, then $\cos(2k\pi/n) \approx \cos(\pi) = 1$. Moreover, $\cos(2n\pi/n) = -1$. If the stochastic matrix $A$ has $B$ (as above) as a principal submatrix, then the vector $v$ may appear as a subvector of the one of the eigenvectors utilised by the Perron cluster approach.

Let $A$ be as above and let $v^{(1)}, \ldots, v^{(m)}$ be eigenvectors associated with eigenvalues near to 1. The PCCA Algorithm attempts to construct a partition

$$\Psi = (\mathcal{E}_1, \ldots, \mathcal{E}_m)$$

of the state space such that if $i$ and $j$ are contained in the same member of $\Psi$, then for all $l$, $|v_i^{(l)} - v_j^{(l)}|$ is relatively small. The vector $v$, above, does not satisfy this property – its entries vary significantly. In a way, such a vector "instructs" the PCCA algorithm to separate the states within a cyclic aggregate. (The algorithm may or may not actually separate these states – the influence of the other selected eigenvectors may overwhelm this incorrect input.)

Paths possess spectra very similar to that of cycles. For example, the eigenvalues of the $n \times n$ stochastic matrix

$$P = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{bmatrix}$$

are

$$\lambda_k = 2\cos\left(\frac{(k-1)\pi}{n-1}\right)$$

for $k = 1, \ldots, n$. When $n$ is large, we see multiple eigenvalues very close to the

eigenvalue 2, implying that the Perron cluster methods are inappropriate for use with a matrix that has a small perturbation of $(1/2)P$ as a principal submatrix.

# Appendix G

# A summary of the properties of the stochastic complement

As before, the results below apply only to discrete-time time-homogeneous Markov chains on finite state spaces.

**The existence of the stochastic complement.** *Let $X$ be a Markov chain on the state space $\mathcal{S}$ with transition matrix $A$; let $\mathcal{C} \subseteq \mathcal{S}$ be a nonempty proper subcollection of $\mathcal{S}$. Then, the following are equivalent:*

1. *the stochastic complement $A \setminus \mathcal{C}$ is defined;*
2. *the collection $\mathcal{C}$ does not contain an entire essential class of states;*
3. *the collection $\mathcal{S} \setminus \mathcal{C}$ contains at least one member from each essential class of states; and*

4. *the principal submatrix $A(\mathcal{C})$ corresponding to $\mathcal{C}$ is properly substochastic.*

*Moreover, $A \setminus \mathcal{C} = I$ if and only if $\mathcal{S} \setminus \mathcal{C}$ contains exactly one member of each essential class of states.*

The above proposition summarises Corollary 2.7 and Propositions 4.2 and 4.11.

**Properties shared by a stochastic matrix and a derived stochastic complement.** *Let $X$ be a Markov chain on the state space $\mathcal{S}$ with transition matrix $A$; let $\mathcal{C} \subseteq \mathcal{S}$ and let $\hat{A} = A \setminus \mathcal{C}$ be the stochastic complement which removes $\mathcal{C}$. The following properties hold:*

1. *The multiplicity of 1 as an eigenvalue of $A$ is equal to the multiplicity of 1 as an eigenvalue of $\hat{A}$.*

2. *The number of distinct essential classes of states with respect to $A$ is equal to the number of distinct essential classes of states with respect to $\hat{A}$.*

3. *If $\mathcal{E}$ is an essential class of states with respect to $A$ then $\mathcal{E} \setminus \mathcal{C}$ is an essential class of states with respect to $\hat{A}$.*

4. *A state $i \in \mathcal{S} \setminus \mathcal{C}$ is recurrent with respect to $A$ if and only if it is recurrent with respect to $\hat{A}$.*

5. *A state $i \in \mathcal{S} \setminus \mathcal{C}$ is transient with respect to $A$ if and only if it is transient with respect to $\hat{A}$.*

6. *For any states $i, j \in \mathcal{S} \setminus \mathcal{C}$, we have $i \preceq j$ with respect to $A$ if and only if $i \preceq j$ with respect to $\hat{A}$.*

7. *Let $\pi$ be a stationary distribution of $A$ and let*

$$\hat{\pi} = \frac{1}{\pi(\mathcal{S} \setminus \mathcal{C})^T \mathbb{1}} \pi(\mathcal{S} \setminus \mathcal{C});$$

*then, the vector $\hat{\pi}$ is a stationary distribution of $\hat{A}$. Moreover, any stationary distribution of $\hat{A}$ can be obtained in this manner.*

*Suppose further that the Markov chain $X$ is reversible. Then, the following additional statements hold:*

8. *The Markov chain associated with $\hat{A}$ is reversible.*

9. *Suppose that $D$ is a positive diagonal matrix such that $DA$ is symmetric. Then, the principal submatrix $\hat{D} = D(\mathcal{S} \setminus \mathcal{C})$ is such that $\hat{D}\hat{A}$ is symmetric.*

10. *Let $\pi$ be a stationary distribution of $A$; then, for all $i, j \in \mathcal{S} \setminus \mathcal{C}$, $\pi_i \hat{a}_{ij} = \pi_j \hat{a}_{ji}$.*

Statements 1 and 2, above, come from Proposition 4.9; statements 3 through 6 come from Proposition 4.8; statement 7 is derived from Proposition 4.6 and Corollary 4.10; statement 8 is Proposition 4.7; and statements 9 and 10 are derived from Propositions 2.12, 4.6 and 4.7 and Corollary 4.10.

# Bibliography

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics, 1994.

[3] Richard A. Brualdi and Herbert J. Ryser. *Combinatorial Matrix Theory*. Cambridge University Press, 1991.

[4] Fan R.K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[5] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 2001.

[6] R.W. Cottle. Manifestations of the Schur complement. *Linear Algebra and its Applications*, 8:189–211, 1974.

[7] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315:39–59, 2000.

[8] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161–184, 2005.

[9] M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25:619–633, 1975.

[10] D. Fritzsche, V. Mehrmann, D.B. Szyld, and E. Virnik. An SVD approach to identifying metastable states of Markov chains. *Electronic Transactions on Numerical Analysis*, 29:46–69, 2008.

[11] J.R. Gilbert, C. Moler, and R. Schreiber. Sparse matrices in MATLAB: design and implementation. *SIAM Review*, 13(1):333–356, 1992.

[12] Chris Godsil and Gordon Royle. *Algebraic Graph Theory*. Springer, 2001.

[13] D.J. Hartfiel and C.D. Meyer. On the structure of stochastic matrices with a subdominant eigenvalue near 1. *Linear Algebra and its Applications*, 272:193–203, 1998.

[14] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[15] M.N. Jacobi. A robust spectral method for finding lumpings and meta stable states of non-reversible Markov chains. *Electronic Transactions on Numerical Analysis*, 37:296–306, 2010.

[16] S. Kirkland. On a question concerning condition numbers for Markov chains. *SIAM Journal on Matrix Analysis and Applications*, 23(4):1109–1119, 2002.

[17] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[18] E. Meerbach, Ch. Schütte, and A. Fischer. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 398:141–160, 2005.

[19] C.D. Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272, 1989.

[20] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):[19 pages], 2006.

[21] James R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[22] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *Journal of Computational Physics*, 151:146–168, 1999. Special Issue on Computational Biophysics.

[23] Eugene Seneta. *Non-negative Matrices and Markov Chains*. Springer-Verlag New York, 1981.

[24] R.M. Tifenbach. On an SVD-based algorithm for identifying meta-stable states of Markov chains. *Electronic Transactions on Numerical Analysis*, 38:17–33, 2011.

[25] Fuzhen Zhang. *The Schur Complement and its Applications*. Springer, 2005.