On the Lumpability of tree-valued Markov Chains

-

Rodrigo Barreto Alves (IME - UERJ)

Joint work with Luiz M. Carvalho (FGV EMAp) and Yuri F. Saporito (FGV EMAp)

September 19, 2025

Outline

Phylogenetic trees

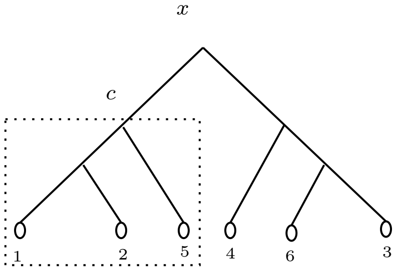


Figure 1: **A tree $x \in \boldsymbol{T}_6$ and one of its clades.** The clade c as a subtree with leaves $\{1, 2, 5\}$ is shown in the dashed rectangle.

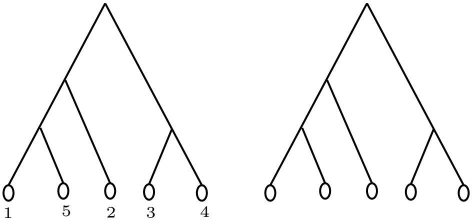


Figure 2: **Labelled and unlabelled rooted trees.**

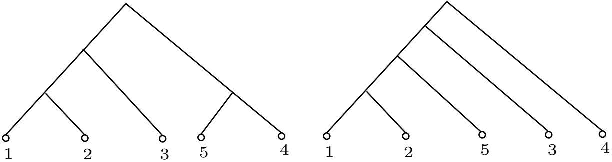


Figure 3: **A balanced (left) and a ladder (right) trees.** These represent the trees with the largest and smallest neighbourhoods in the rSPR graph, respectively.

Tree space and Subtree prune-and-regraft (SPR)

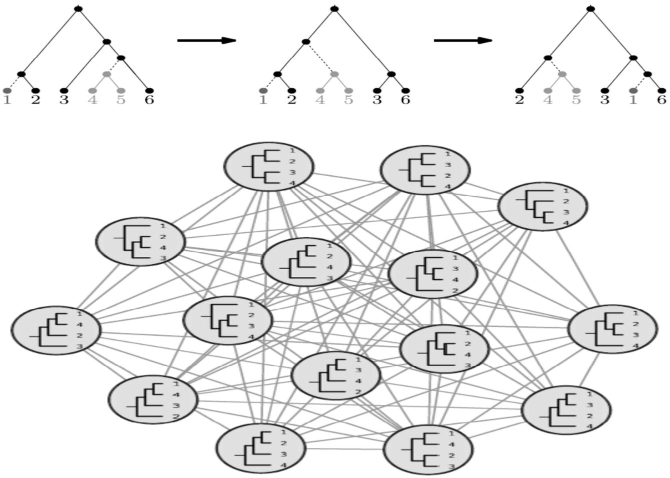


Figure 4: Figure from Whidden and Matsen IV (2017). For simplicity $d_{r-SPR}(x,y) := d(x,y)$.

Properties of the rSPR-graph

- **Diameter:** The diameter of the rSPR-graph is $O(n)$ (Song, 2003).
- **Connectivity:** From the proof of the diameter, it can be inferred that the rSPR-graph is connected.
- **Neighbourhood size:** Each tree has $O(n^2)$ neighbours.
  - Maximum neighbours: balanced tree, $4(n-2)^2 - 2\sum_{j=1}^{n-2} \lfloor \log_2(j+1) \rfloor$
  - Minimum neighbours: ladder tree. $3n^2 - 13n + 14$.
- **Complexity:** Computing the distance between two trees is NP-hard (Bordewich and Semple, 2005).

Motivation

If one has a Markov process $(X_k)_{k\geq 0}$ on the space of trees, is the induced process $(Y_k(c))_{k\geq 0} \in \{0,1\}$ for each clade c also Markov?

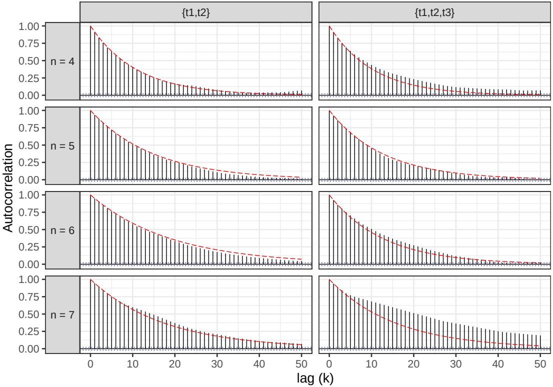


Figure 5: Autocorrelation spectra of clade indicators for the LMH. We show the empirical autocorrelation spectra up to lag $k = 50$ for indicators of clades when sampling from a LMH with $\rho = 0.9$ on a single realisation.

SPR Metropolis-Hastings random walk

We define a Metropolis Hastings random walk on a SPR graph

$$p_{\mathrm{MH}}(x,y)=\begin{cases}\frac{1}{|N(x)|}\min\left\{1,\frac{|N(x)|}{|N(y)|}\right\},y\in N(x),\\ 1-\sum_{z\in N(x)}\frac{1}{|N(x)|}\min\left\{1,\frac{|N(x)|}{|N(z)|}\right\},y=x\\ 0,y\notin N(x).\end{cases}$$

Where $N(x):=\{y\in \boldsymbol{T}_n : d(x,y)=1\}$.

This leads to a uniform distribution, i.e., $\pi_{\mathrm{MH}}(t)=1/|\boldsymbol{T}_n|$ for all $t\in\boldsymbol{T}_n$.

Lumpability

Let $(X_k)_{k\geq 0}$ be Markov chain on $\mathcal{S} = \{f_1, f_2, \ldots, f_r\}$, $\bar{S} = \{E_1, E_2, \ldots, E_v\}$ a partition of $\mathcal{S}$ and $(Y_k)_{k\geq 0}$ is the projected process on $\bar{S}$.

Definition

Let $(X_k)_{k\geq 0}$ be Markov chain on a finite state-space $\mathcal{S} = \{f_1, f_2, \ldots, f_r\}$ with initial distribution μ_0 and transition probabilities matrix $\boldsymbol{P}$. We say $(X_k)_{k\geq 0}$ is **lumpable** with respect to a partition of $\bar{S} = \{E_1, E_2, \ldots, E_v\}$ of the state space if the projected process $(Y_k)_{k\geq 0}$ on $\bar{S}$ is also a Markov chain for any μ_0.

If $(X_k)_{k\geq 0}$ is lumpable with respect to the partition $\bar{S}$ for any $x, y \in E_i$ we have

$$\sum_{z\in E_j} p(x,z) = \sum_{z\in E_j} p(y,z)\,.$$

Shape-lumpability of tree-valued Markov chains

- Let $\bar{F} := \{F_1, F_2, \dots, F_v\}$ be a tree shape partition of $\boldsymbol{T}_n$.
- It is important to notice that if $x, y \in F_i$, we have $|N(x)| = |N(y)|$.
- For a $x \in \boldsymbol{T}_n$, we define

$$F_j^x := \{y \in T_n : d(x,y) = 1 \text{ and } y \in F_j\} = N(x) \cap F_j \,.$$

Lemma

Let x and y be trees in $\boldsymbol{T}_n$, such that they have the same shape, i.e. $x, y \in F_i$. Then for all $j \in \{1, 2, \dots, v\}$ we have $|F_j^x| = |F_j^y|$.

Shape-lumpability of tree-valued Markov chains

Theorem

Consider the SPR Metropolis-Hastings random walk. Let $\bar{F} := \{F_1, F_2, \ldots, F_v\}$ be a tree shape partition of $\boldsymbol{T}_n$. Then we have that the SPR Metropolis-Hastings random walk is lumpable with respect to the partition $\bar{F} := \{F_1, F_2, \ldots, F_v\}$.

Idea of the proof of the Theorem (MH)

Fix a $j \in \{1, 2, \ldots, v\}$ and $i \in \{1, 2, \ldots, v\}$. For all $x, y \in F_i$, we have

$$\sum_{z \in F_j} p(x,z) - \sum_{z \in F_j} p(y,z) =$$

$$\sum_{z \in F_j^x} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(z)|}\right\} - \sum_{z \in F_j^y} \frac{1}{|N(y)|} \min\left\{1, \frac{|N(y)|}{|N(z)|}\right\} .$$

For $z' \in F_j^x$. We have two possible situations $|N(x)| \geq |N(z')|$ or $|N(x)| < |N(z')|$.

Lumping error and ϵ-lumpability

Definition

Consider again a partition $\bar{S} = \{E_1, \ldots, E_v\}$ of $\mathcal{S}$. For $x, y \in E_i$, define the **lumping error** as

$$R_{i,j}(x,y) = \sum_{z \in E_j} p(x,z) - \sum_{z \in E_j} p(y,z).$$

When $|R_{i,j}(x,y)| \leq \epsilon$ for every pair x, y, we say the Markov chain is ϵ-almost lumpable with respect to $\bar{S}$.

Clade partition of tree-space

Let $\boldsymbol{C}_n$ be the space of all partitions of the leaf set (i.e. $\boldsymbol{C}_n$ is the space of all clades). Denote $C(x)$ as the set of clades that compose $x \in \boldsymbol{T}_n$.

Definition

Let $\bar{S}_n(c) = \{S_0(c), S_1(c)\}$ be the partition of $\boldsymbol{T}_n$ induced by clade $c \in \boldsymbol{C}_n$, for which we will write $S_0(c) := \{y \in \boldsymbol{T}_n : c \notin C(y)\}$ and $S_1(c) := \{y \in \boldsymbol{T}_n : c \in C(y)\} = \boldsymbol{T}_n \setminus S_0(c)$.

- For $x \in S_1(c)$ we set $A_1^{x,c} := S_1(c) \cap N(x)$ and $A_0^{x,c} := S_0(c) \cap N(x)$.
- For $x \in S_0(c)$, we denote $B_1^{x,c} := N(x) \cap S_1(c)$ and $B_0^{x,c} := N(x) \cap S_0(c)$.

ε-Lumpability for the Metropolis-Hastings random walk

Theorem

Consider the SPR Metropolis-Hastings random walk $(X_k)_{k\geq 0}$ and the partition $\bar{S} := \{S_0(c), S_1(c)\}$ of $\boldsymbol{T}_n$. Then the lumping error for $(X_k)_{k\geq 0}$ with respect to the partition $\bar{S}$ is evaluated, for $|c| = 2$,

$$\varepsilon = \begin{cases} \varepsilon(S_0(c), S_1(c)), & \text{for } 4 \leq n \leq 8, \\ \varepsilon(S_1(c), S_0(c)), & \text{for } n \geq 9\,, \end{cases}$$

Now for $3 \leq |c| \leq \lfloor n^{1/2} \rfloor$ and $n \geq 9$, $\varepsilon = \varepsilon(S_0(c), S_1(c))$.

ε-Lumpability for the Metropolis-Hastings random walk

For simplicity $S_0(c) := S_0$ and $S_1(c) := S_1$.

For all $x, y \in S_1$, by Lemma 5.5 in Whidden and Matsen IV (2017) we have

$$
\begin{aligned}
R_{S_1,S_0}(x,y) &= \sum_{z \in A_0^{x,c}} p(x,z) - \sum_{z \in A_0^{y,c}} p(y,z) \\
&= \sum_{z \in A_0^{x,c}} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(z)|}\right\} - \sum_{z \in A_0^{y,c}} \frac{1}{|N(y)|} \min\left\{1, \frac{|N(y)|}{|N(z)|}\right\} \\
&\leq \frac{|A_0^{x,c}|}{|N(x)|} - \frac{5|A_0^{y,c}|}{6|N(y)|} = \varepsilon(S_1, S_0)\,.
\end{aligned}
$$

ε-Lumpability for the Metropolis-Hastings random walk

For all $x, y \in S_0$, by Lemma 5.5 in Whidden and Matsen IV (2017) we have

$$
\begin{aligned}
R_{S_0(c),S_1(c)}(x,y) &= \sum_{z \in B_1^{x,c}} p(x,z) - \sum_{z \in B_1^{y,c}} p(y,z) \\
&\leq \sum_{z \in B_1^{x,c}} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(z)|}\right\} - \sum_{z \in B_1^{y,c}} \frac{1}{|N(y)|} \min\left\{1, \frac{|N(y)|}{|N(z)|}\right\} \\
&\leq \frac{|B_1^{x,c}|}{3n^2 - 13n + 14} - \frac{5|B_1^{y,c}|}{6|N(y)|} = \varepsilon(S_0, S_1)\,.
\end{aligned}
$$

Lumpability Experiment

- We consider for the original process $(X_k)_{k\geq 0}$ on $\boldsymbol{T}_n$,

$$P_{MH}^{\rho} = \rho I + (1-\rho) P_{MH}\,,$$

  for $\rho \in \{0, 0.1, 0.5, 0.9\}$.

- We generate the Lumped transition probability matrix, Q, for the projected process $(Y_k)_{k\geq 0}$ on $\bar{F} = \{F_1, F_2, \ldots, F_v\}$.

$$q(F_i, F_j) = p(x, F_j)\,,$$

  where $x \in F_i$ and $i, j \in \{1, 2, \ldots, v\}$.

- We ran 500 independent replicates of each chain with 10000 iterations each. Define $\hat{\mu}_k^j$ and $\hat{\nu}_k^j$ as the empirical measures of $(X_k^j)_{k\geq 0}$ and $(Y_k^j)_{k\geq 0}$.

Lumpability Experiment

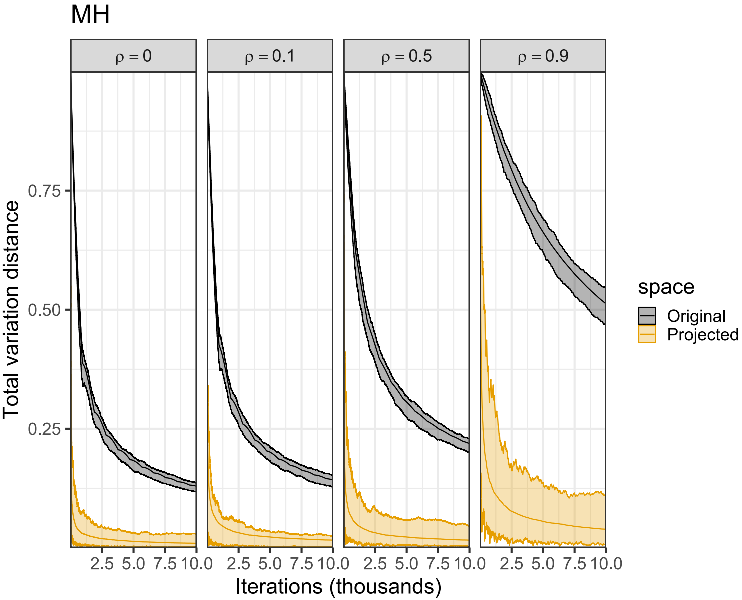
- We generate a new empirical measure from $\hat{\nu}_k^j$, assigning it the domain $\boldsymbol{T}_n$ and denoting it as $\hat{\eta}_k^j$. we compute $p_k^{F_i} := \hat{\nu}_k^j(F_i) \times |F_i|^{-1}$ and for each tree $x \in F_i$, we assign $\hat{\eta}_k^j(x) = p_k^{F_i}$.

- Measuring the distance to the stationary distribution

$$
\begin{aligned}
m_k &:= \min_{j \in \{1,2,\ldots,500\}} ||\hat{\mu}_k^j - \pi_X|| \\
M_k &:= \max_{j \in \{1,2,\ldots,500\}} ||\hat{\mu}_k^j - \pi_X|| \\
E_k &:= \frac{1}{500} \sum_{j=1}^{500} ||\hat{\mu}_k^j - \pi_X|| \,.
\end{aligned}
$$

The same metrics will be calculated for $\hat{\eta}_k^j$ at each iteration.

Results for $n = 6$



ε-Lumpability Experiment for $|c| = 2$

- We define $\eta_X := (\pi_X(S_0(c)), \pi_X(S_1(c)))$ where $(X_k)_{k\geq 0}$ is a MH on $\boldsymbol{T}_n$. Define $(Y_k)_{k\geq 0}$ as the projected process on the partition $\bar{S} = \{S_0(c), S_1(c)\}$.
- Define $(\tilde{Y}_k)_{k\geq 0}$ as the auxiliary process on $\{\tilde{S}_0^c, \tilde{S}_1^c\}$ and $\hat{\mu}_{\tilde{Y}_k}$ the empirical measure.

$$||\hat{\mu}_{\tilde{Y}_k} - \eta_X|| \leq ||\hat{\mu}_{\tilde{Y}_k} - \hat{\mu}_{Y_k}|| + ||\hat{\mu}_{Y_k} - \eta_X||\,.$$

- The auxiliary Markov Chain

$$\tilde{p}(S_0^c, S_1^c) = \tilde{p}(S_1^c, S_1^c) = \frac{1}{2}\Big(\frac{2n-5}{3n^2-13n+14} + \frac{5}{6(4(n-2)^2 - 2\sum_{j=1}^{n-2}\lfloor \log_2(j+1)\rfloor)}\Big).$$

ε-Lumpability Experiment for $|c| = 2$

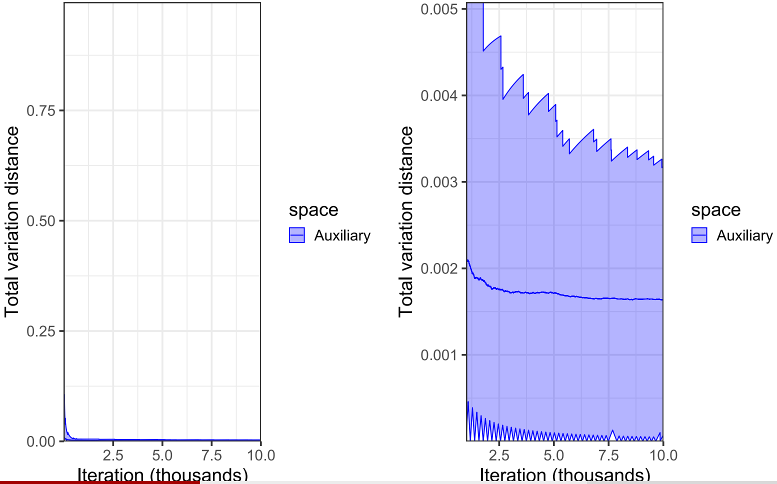
- We ran 500 independent replicates of $(\tilde{Y}_k^j)_{k\geq 0}$, the auxiliary one, with 10.000 iterations each. Denote $\hat{\mu}_{\tilde{Y}_k^j}$ as the empirical measure.
- Measuring the distance to the stationary distribution

$$
\begin{aligned}
m_k &:= \min_{j\in\{1,2,\ldots,500\}} ||\hat{\mu}_{\tilde{Y}_k^j} - \pi_X|| \\
M_k &:= \max_{j\in\{1,2,\ldots,500\}} ||\hat{\mu}_{\tilde{Y}_k^j} - \pi_X|| \\
E_k &:= \frac{1}{500}\sum_{j=1}^{500} ||\hat{\mu}_{\tilde{Y}_k^j} - \pi_X|| \,.
\end{aligned}
$$

Results for $|c| = 2$ and $n = 100$

$$\eta_X = (0.9949, 0.0051),$$

MH with n = 100



Open problems

- Develop methods to construct auxiliary processes on smaller spaces to improve Monte Carlo estimation.
- Find partitions that minimize lumping error while retaining interpretability.
- Explore the relationship between lumping error and convergence speed in projected processes.
- Investigate how to generalize the findings to more realistic posterior distributions and target processes.

Idea of the proof that of the Lemma

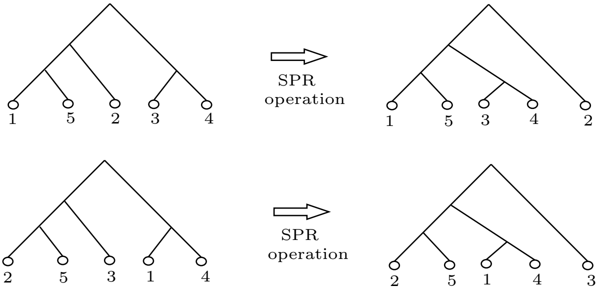


Figure 6: Same SPR operation on different trees, however with the same shape.

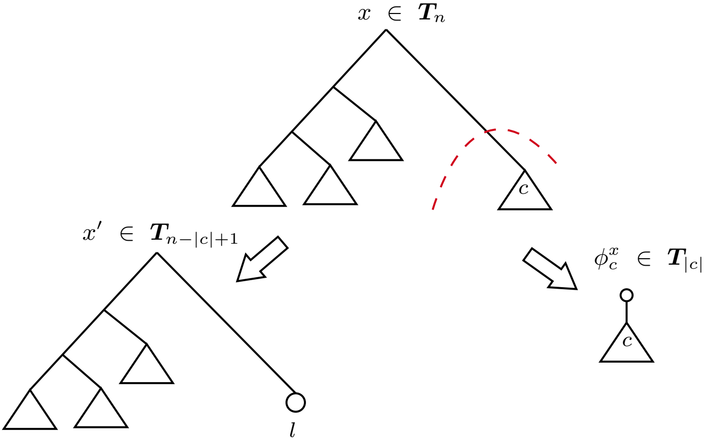
Some results about clade partition of tree-space

- For a tree $x \in S_1(c)$ we set $A_1^{x,c} := S_1(c) \cap N(x)$ and $A_0^{x,c} := S_0(c) \cap N(x)$.
- We define $f_c : S_1(c) \to \boldsymbol{T}_k$, where $k := n - |c| + 1$.

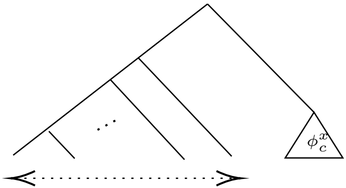
Lemma

If $x \in S_1(c)$ then $|A_1^{x,c}| = |N(x')| + |N(\phi_c^x)|$, where $x' = f_c(x)$.

Idea of the proof that $|A_1^{x,c}| = |N(x')| + |N(\phi_c^x)|$.



Some results about clade partition



Lemma

Let $x \in S_1(c)$ be such in Figure above. Then for all $w \in S_1(c)$ we have $|A_0^{x,c}|/|N(x)| \geq |A_0^{w,c}|/|N(w)|$ and

$$\frac{|A_0^{x,c}|}{|N(x)|} = \frac{-8|c|^2 + 8|c|n + 6|c| - 8n - 2}{3n^2 - 2|c|^2 + 2|c|n - 15n + 16} \quad \textit{for } |c| \geq 3 \;, \textit{and}$$

$$\frac{|A_0^{x,c}|}{|N(x)|} = \frac{8n - 22}{3n^2 - 11n + 8} \quad \textit{for } |c| = 2\,.$$

Some results about Clade partition of tree-space

Let $x \in S_0(c)$, we denote $B_1^{x,c} := N(x) \cap S_1(c)$. Let $y \in S_0(c)$, we define I as a set of leaves such that we have a subtree ϕ_I and $c \cup I$ generates a subtree.

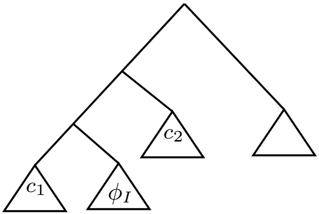


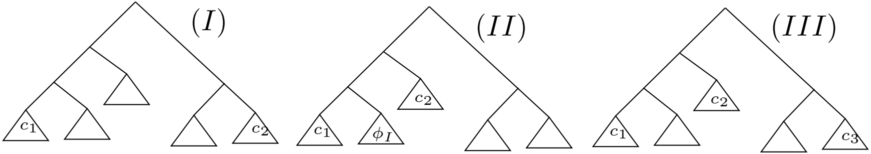
Figure 7: We have c_1 and c_2 clades such that $c_1 \cap c_2 = \emptyset$ and $c_1 \cup c_2 = c$.

Some results about Clade partition of tree-space

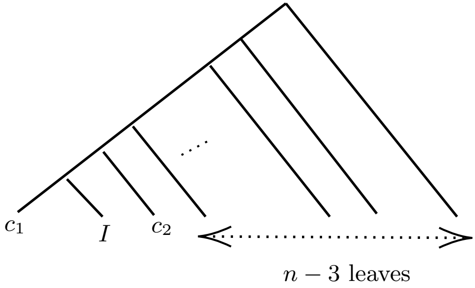
Lemma

Let $x \in S_0(c)$ *and* $3 \leq |c| \leq n-3$ *then*

$$|B_1^{x,c}| = \begin{cases} 2(|c|-1), & \textit{(I)} \\ 2(n-|I|)-3, & \textit{(II)} \\ 0 \quad \textit{(III)}. \end{cases}$$



Some results about Clade partition of tree-space



Lemma

Let $x \in S_0(c)$ *be such in Figure above and* $c := \{c_1, c_2\}$*. Then for all* $w \in S_0(c)$ *we have* $|B_1^{x,c}|/|N(x)| \geq |B_1^{w,c}|/|N(w)|$ *and*

$$\frac{|B_1^{x,c}|}{|N(x)|} = \frac{2n-5}{3n^2 - 13n + 14}.$$

ε-Lumpability for the Metropolis-Hastings random walk

Remembering the notation

Consider a Markov process on a state space $\mathcal{S}$ with a partition $\bar{S} = \{E_1, \dots, E_h\}$. Then for any $x, y \in E_i$ we have

$$|R_{E_i,E_j}(x,y)| = \Big| \sum_{z \in E_j} p(x,z) - p(y,z) \Big| \leq \varepsilon(E_i, E_j)\,.$$

The lumping error ε will be such that $\varepsilon(E_i, E_j) \leq \varepsilon$ for all $i, j \in \{1, 2, \dots, h\}$.