

Research Article

Statistical Identification of Markov Chain on Trees

Xuyan Xiang^{1,2,3}, Xiao Zhang² and Xiaoyun Mo⁴

¹Hunan Province Cooperative Innovation Center for the Construction and Development of Dongting Lake Ecological Economic Zone, School of Mathematics and Computational Science, Hunan University of Arts and Science, Changde 415000, China

²School of Mathematics and Statistics, Jishou University, Jishou 416000, China

³HPCSIP, Ministry of Education of China, Hunan Normal University, Changsha 410081, China

⁴School of Mathematics and Statistics, Hunan University of Finance and Economics, Changsha 410205, China

Correspondence should be addressed to Xiaoyun Mo; moxyun72@163.com

Received 8 October 2017; Accepted 13 February 2018; Published 28 March 2018

Academic Editor: Pasquale Palumbo

Copyright © 2018 Xuyan Xiang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The theoretical study of continuous-time homogeneous Markov chains is usually based on a natural assumption of a known transition rate matrix (TRM). However, the TRM of a Markov chain in realistic systems might be unknown and might even need to be identified by partially observable data. Thus, an issue on how to identify the TRM of the underlying Markov chain by partially observable information is derived from the great significance in applications. That is what we call the statistical identification of Markov chain. The Markov chain inversion approach has been derived for basic Markov chains by partial observation at few states. In the current letter, a more extensive class of Markov chain on trees is investigated. Firstly, a type of a more operable derivative constraint is developed. Then, it is shown that all Markov chains on trees can be identified only by such derivative constraints of univariate distributions of sojourn time and/or hitting time at a few states. A numerical example is included to demonstrate the correctness of the proposed algorithms.

1. Introduction

The classic study of continuous-time homogeneous Markov chains usually has a natural assumption that it has a known transition probability matrix or a transition rate matrix (TRM). However, the TRM of a Markov chain in realistic systems might be unknown although it is an underlying one (named as the underlying Markov chain), and that is what needs to be identified by observable data. For a Markov chain with a finite state space, it is determinative provided that its TRM is identified. In some realistic systems, moreover, one probably observes the partial motion of the underlying Markov chain. For example, one can only observe the sojourn sequences and hitting sequences on one of the states. Subsequently, one can fit the PDF of sojourn (resp., hitting) time by statistical techniques. It is easy and trivial to identify the TRM of this Markov chain if we know exactly the PDFs of sojourn time and hitting time of every state.

A natural question arises as to whether it is possible to identify the TRM by those few states. Due to the transition relations of states which intercommunicate in the underlying

Markov chain, it is indeed possible. Thus, an issue on how to identify the TRM of the underlying Markov chain by partially observable information is derived from the great significance in applications.

In fact, the continuous-time homogeneous Markov model which described the gating kinetics of single ion channel is a good example to show the real application in biophysics and neuroscience; for example, see [1, 2] and further publications. In experiments, the transition between the various states cannot be directly observed and only transitions between a small number of open states are observable. It is straightforward to derive the sojourn time and hitting time distributions for a single state or a small number of states. The PDFs of sojourn or hitting time generally take the form of a sum of exponentials (e.g., [2–4]). And the best-fitting approach and function can be found with a number of readily available computer programs [5–7].

Unfortunately, once the fitting is completed, it is very difficult to use these PDFs to identify the TRM of the underlying Markov chain. The most difficult challenge lies in finding out the hidden solutions and algorithms to perform the

reverse operation. Thus, the identification has always been addressed directly using the maximum likelihood estimate (MLE) (e.g., see [8–11] and further references) rather than performing this inversion. Another series of publications using canonical forms of reduced dimensions analyze finite two-state trajectories, in which the system is aggregated into only two states, defined as the *on* state and the *off* state (irreversible transitions are also allowed); see [12, 13] and the references therein for a review. All of them are powerful. However, each approach has pros and cons. None of them can perfectly identify all kinds of Markov chains. In general, for example, the sojourn times f_C and f_O may not be sufficient for parameterizing a given Markov model. This is shown by [14] where, in general, f_C and f_O as well as bivariate distributions of subsequent open and closed times, f_{OC} and f_{CO} , are necessary.

Note that those discarded the prior information from the underlying topology. Thus, it is indeed possible to perform the complicated inversion since the prior information from the underlying architecture of Markov chain is very helpful to accomplish it. This is what we call *statistical identification of Markov chain* (SIMC). We developed a Markov chain inversion approach (MCIA) to open the possibility of solving the SIMC. This issue can be reduced to three steps for applications. The first one is to obtain the necessary data at the fewest possible states by preprocessing the observed data. The second one is to accurately fit the corresponding PDFs of states required by the next step. The final one is to find out the algorithm to identify the TRM of the underlying Markov chain by those PDFs. Throughout this paper, *we focus our attention on the final one* since there are a number of readily available approaches to fit those PDFs.

We explore a class of Markov models where reversibility is maintained. Since some realistic systems obey the principle of microscopic reversibility or detailed balance, this implied the reversibility of the Markov chain. At least in applications to ion channels, an assumption of reversibility is plausible in many circumstances, corresponding to channels which are not coupled to a free energy source such as an ion concentration gradient; for discussion of this assumption, see [15]. As an example, [16, 17] derived a criterion to identify whether it is reversible or not in a three-state Markov system based on the coarse-grained information of two-state trajectories. Throughout the rest of this paper, the unqualified term “Markov chain” usually refers to a reversible Markov chain.

It is well known that classic architectures include star-graph, linear, and cycle chains. Thus, some general constraints and the corresponding algorithms for them were explored in former publications [18–20]. There is no obvious strategy for selecting the constraints to use for a general model. A key task is to find out the feasible solutions and the corresponding algorithms to different classes. Then, a subsequent question is whether more complicated architectures are identified by such approach. On this point, any one of them can be viewed as a tree provided that it contains no cycles. This implies that the tree is a very important extension of the star graph and line graph and represents one of the two classes of topologies (whether containing cycles or not). It

is known that all Markov chains on trees are reversible [21]. Thus, SIMC on trees will be exploited in this letter, although some of them are still not applied in real systems. It is derived that all Markov chains on trees are identified by our approach. It is mentioned that, for SIMC on trees, the greatest contribution should be to discover the fundamental way and the corresponding algorithm to identify the TRM due to the challenge of performing the reverse operation as stated earlier. Hence, the work further opens up the possibility of carrying out the statistical identification for all reversible Markov chains.

Such approach has obvious advantages over MLE in that it can identify the most reversible Markov chain without the requirement of bivariate distributions of subsequent sojourn time and hitting time and in that the computation is accurate based on the accurate sojourn time PDFs and the prior information about the underlying topological structure of the Markov chain.

This paper is organized as follows. After recalling the sojourn time distribution, a type of a more operable derivative constraint is developed in Section 2. In Section 3, the solution to SIMC on trees is provided by such derivative constraints. In addition, we address the solutions to three particular and regular representatives of trees, including theoretical conclusions, proofs, and algorithms. Finally, a numerical example is included to demonstrate the correctness of the proposed model.

2. Distributions of Hitting Time and Its Derivative Relationships

Throughout this section, $\text{diag}(\cdots)$ denotes a diagonal matrix whose entries on the diagonal are those of them, and \top denotes transpose, $\mathbf{1}$ a column vector of ones, and $\mathbf{0}$ a matrix (vector) of zeros, with the dimensions of these being clear from their context. Furthermore, A_{ij} and A_i denote the (i, j) th entry and the i th column vector of the matrix A , respectively.

Let $\{X_t : t \geq 0\}$ be a reversible Markov chain with finite state space $\mathbf{S} = \{0, 1, \dots, M\}$, conservative TRM $Q = (q_{ij})_{S \times S}$ such that $Q\mathbf{1} = \mathbf{0}$ ($q_{ij} \geq 0$, $j \neq i$; $q_i \equiv -q_{ii} > 0$), and equilibrium distribution $\tilde{\pi} = [\pi_0, \pi_1, \dots, \pi_M]^\top$ such that

$$\begin{aligned} \Pi^{1/2} Q \Pi^{-1/2} &= \left(\Pi^{1/2} Q \Pi^{-1/2} \right)^\top, \\ \tilde{\pi}^\top \mathbf{1} &= 1, \end{aligned} \quad (1)$$

where $\Pi = \text{diag}(\pi_0, \pi_1, \dots, \pi_M)$, and the first one (i.e., $\pi_i q_{ij} = \pi_j q_{ji}$ for any $i, j \in S$) implies the reversibility.

Assume that the state space may be partitioned as $\mathbf{S} = \mathcal{C} \cup \mathcal{O}_1 \cup \mathcal{O}_2 \cup \cdots \cup \mathcal{O}_s$ into $r + s (\geq 1)$ classes, where \mathcal{C} denotes the class of closed states (unobservable), $\mathcal{O}_1 = \{1, 2, \dots, r\}$ the class of open states being distinguishable (i.e., having r different open levels), and \mathcal{O}_i ($i = 2, \dots, s$) the class of aggregated open states (i.e., having other common open levels); for example, see [2].

A proper subset \mathbf{O} of the open class $\mathcal{O} = \mathcal{O}_1 \cup \mathcal{O}_2 \cup \cdots \cup \mathcal{O}_s$ is observed such that either $\mathcal{O}_k \subseteq \mathbf{O}$ or $\mathcal{O}_k \cap \mathbf{O} = \emptyset$ for $k = 2, \dots, s$ (this condition allows one to obtain the sojourn

time and hitting time sequences of \mathbf{O} from experimental recording). Let $\mathbf{C} = \mathbf{S} - \mathbf{O}$. The state space is now partitioned as $\mathbf{S} = \mathbf{O} \cup \mathbf{C}$. The TRM Q may be partitioned as, by the classes \mathbf{O} and \mathbf{C} ,

$$Q = \begin{pmatrix} Q_{oo} & Q_{oc} \\ Q_{co} & Q_{cc} \end{pmatrix}. \quad (2)$$

The corresponding partition is used for the equilibrium distribution $\tilde{\pi}$; that is, $\tilde{\pi}^\top = [\tilde{\pi}_o^\top, \tilde{\pi}_c^\top]$.

2.1. Distributions of Hitting Time. The sojourn time and hitting time of \mathbf{O} can be defined as $\sigma = \inf\{t > 0, X_t \notin \mathbf{O}\}$ and $\tau = \inf\{t > 0, X_t \in \mathbf{O}\}$, with the usual convention that the infimum over an empty set is infinity, respectively. They are also the hitting time and sojourn time of \mathbf{C} , respectively.

For some basic results on aggregated continuous-time Markov chains, refer to [1, 2, 20, 22]. Thus, just some of them, required in the sequel, are now recalled.

For ease of exposition, let N be the number of \mathbf{C} , and $\mathbf{C} \equiv \{1, 2, \dots, N\}$, $\Pi_c \equiv \text{diag}(\pi_1, \pi_2, \dots, \pi_N)$. Note that Q_{cc} is nonsingular since $\{X(t)\}$ is irreducible, and hence all the eigenvalues of Q_{cc} have strictly negative real parts. Assume that $-\alpha_1, -\alpha_2, \dots, -\alpha_N$ are all real eigenvalues of Q_{cc} such that $\alpha_i > 0$. By symmetric reparameterization with respect to $(\cdot, \cdot)_{\Pi_c}$, Q_{cc} can be diagonalized with an orthogonal matrix E and we can obtain a matrix $W = \Pi_c^{-1/2} E$ and then set $D = W^{-1}$ and $A = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_N)$. Then, one can get

$$\begin{aligned} Q_{cc} &= -D^{-1}AD, \\ D^\top D &= \Pi_c, \\ DQ_{cc} &= -AD, \\ \Pi_c Q_{cc} &= -D^\top AD. \end{aligned} \quad (3)$$

Theorem 1. *The PDF of hitting time of \mathbf{O} (resp., the sojourn time of \mathbf{C}) takes the following form:*

$$\begin{aligned} f_\tau(t) &= (\tilde{\pi}_c^\top \mathbf{1})^{-1} (D\mathbf{1})^\top A \exp(-At) D\mathbf{1} \\ &= \sum_{i=1}^N \gamma_i \exp(-\alpha_i t), \quad t > 0. \end{aligned} \quad (4)$$

Note that the above is a sum of N -exponentials, $\gamma_i \exp(-\alpha_i t)$, where $-\alpha_i$ ($\alpha_i > 0$) are the real eigenvalues of Q_{cc} and $\gamma_i = \alpha_i (D\mathbf{1})_i^\top (D\mathbf{1})_i (\tilde{\pi}_c^\top \mathbf{1})^{-1}$.

In particular, if \mathbf{O} only contains a single state i , that is, $\mathbf{O} = \{i\}$, then the PDF of the sojourn time degenerates to an exponential form.

Corollary 2. *The PDF of the sojourn time σ at a single state i takes a form of an exponential*

$$f_\sigma(t) = q_i \exp(-q_i t), \quad t > 0. \quad (5)$$

2.2. The Derivative Constraints. Let $\Gamma = (\gamma_1, \dots, \gamma_N)$ and

$$d_n = \Gamma A^{n-1} \mathbf{1} = \sum_{i=1}^N \gamma_i \alpha_i^{n-1}, \quad n \geq 1, \quad (6)$$

$$c_n = (\pi_c^\top \mathbf{1}) d_n = (1 - \pi_o^\top \mathbf{1}) d_n.$$

Then, $(-1)^n d_n$ is the n th derivative of the hitting time PDF of \mathbf{O} at $t = 0$.

By (6), $d_n = (\pi_c^\top \mathbf{1})^{-1} \sum_{i=1}^N (D\mathbf{1})_i^\top (D\mathbf{1})_i \alpha_i^n = (\pi_c^\top \mathbf{1})^{-1} (D\mathbf{1})^\top A^n (D\mathbf{1})$. By (3), $DQ_{cc} = -AD$, and then $D^\top A^2 D = -D^\top ADQ_{cc} = (-1)^2 \Pi_c Q_{cc}^2$. One can apply such operation repeatedly to conclude $D^\top A^n D = (-1)^n \Pi_c Q_{cc}^n$.

A class of constraints commonly used can be derived from the derivatives of the hitting time distribution at $t = 0$.

Theorem 3. *The derivatives of the hitting time distribution of \mathbf{O} at $t = 0$ are related to the transition rates with the formula*

$$c_n = (-1)^n \mathbf{1}^\top \Pi_c Q_{cc}^n \mathbf{1} = (-1)^n \sum_{i \in \mathbf{C}} \pi_i \sum_{j \in \mathbf{C}} (Q_{cc}^n)_{ij}. \quad (7)$$

It is noted that many other constraints are required to identify the general Markov chains [20]. However, derivative constraints are the most common ones. Thus, we will more deeply discover the intrinsic laws in derivative constraints. Further, we will show that only by applying the derivative constraints is it possible to solve most of the Markov chains such as the trees; see the next section.

In particular, the most useful and commonly used constraints are the first three ones.

Corollary 4. *The following equations hold:*

$$\begin{aligned} c_1 &= \sum_{j \in \mathbf{C}} \pi_j \sum_{i \in \mathbf{O}} q_{ji}, \\ c_2 &= \sum_{j \in \mathbf{C}} \pi_j \left(\sum_{i \in \mathbf{O}} q_{ji} \right)^2, \end{aligned} \quad (8)$$

$$c_3 = - \sum_{s \in \mathbf{C}} \sum_{j \in \mathbf{C}} \pi_s q_{sj} \left(\sum_{l \in \mathbf{O}} q_{sl} \sum_{i \in \mathbf{O}} q_{ji} \right).$$

In particular, a single state i has a more simple and visual form of c_n , and by far the most common and convenient are those for a single state.

Corollary 5. *Observing a single state i , one can get the following equations for a tree:*

$$\begin{aligned} c_1 &= \sum_{j \neq i} \pi_j q_{ji}, \\ c_2 &= \sum_{j \neq i} \pi_j (q_{ji})^2, \\ c_3 &= \sum_{j \neq i} \pi_j (q_{ji})^2 q_j, \end{aligned}$$

$$\begin{aligned}
c_4 &= \sum_{j \neq i} \pi_j (q_{ji})^2 \left[\sum_{s \in \mathbf{C}} q_{js} q_{sj} \right], \\
c_{2s+1} &= \sum_{j \neq i} \pi_j (q_{ji})^2 \left[\sum_{a_1, \dots, a_{s-1} \in \mathbf{C}} q_{ja_1} q_{a_1 a_2} \right. \\
&\quad \left. \cdots q_{a_{s-2} a_{s-1}} q_{a_{s-1} a_s} \cdots q_{a_2 a_1} q_{a_1 j} \right], \\
c_{2s+2} &= \sum_{j \neq i} \pi_j (q_{ji})^2 \left[\sum_{a_1, \dots, a_s \in \mathbf{C}} q_{ja_1} q_{a_1 a_2} \cdots q_{a_{s-1} a_s} q_{a_s a_{s-1}} \right. \\
&\quad \left. \cdots q_{a_2 a_1} q_{a_1 j} \right].
\end{aligned} \tag{9}$$

The right-hand side of the first equation gives a sum over all exit rate flows for state i because of $\pi_j q_{ji} = \pi_i q_{ij}$. Conditional upon the first equation, the second and the third one are obvious. The expression within the bracket of c_{2s+1} (resp. c_{2s+2}) is the sum over rates of possible $(2s-1)$ -step (resp., $2s$ -step) transitions departing from a state j , having a direct transition to i , and returning to itself. Those possible transitions imply the intrinsic relationships to Q_{cc} so that most transitions can be identified by those hitting PDFs.

By the end of this section, we show that the above c_n ($n \geq 1$) are known from the required PDFs at \mathbf{O} and follow from Lemma 6. First of all, Corollary 2 shows that $q_i = 1/E\sigma$, where $E\sigma$ is the expected sojourn time for state i . Corollary 5 shows that $c_1 = \sum_{j \neq i} \pi_j q_{ji} = \sum_{j \neq i} \pi_i q_{ij} = \pi_i q_i$. In addition, $c_1 = (1 - \pi_i)d_1$. It suffices to yield $\pi_i = d_1/(q_i + d_1) = c_1 E\sigma$, which implies the following lemma.

Lemma 6. *The PDFs of sojourn time and hitting time at a single state i show that*

$$\begin{aligned}
q_i &= \frac{1}{E\sigma}, \\
\pi_i &= \frac{d_1}{(q_i + d_1)} = c_1 E\sigma.
\end{aligned} \tag{10}$$

For observations at \mathbf{O} , one can obtain the corresponding d_n . Thus, (6) shows that c_n ($n \geq 1$) is known, which is equal to d_n times a known constant $\pi_c \mathbf{1}$ from the observation at \mathbf{O} . The reason is as follows. Lemma 6 shows that, based on statistics for a single state i , one can obtain by its PDF π_i for any $i \in \mathbf{O}$, followed by $c_n = \pi_c^\top \mathbf{1} d_n = (1 - \sum_{i \in \mathbf{O}} \pi_i) d_n$ ($n \geq 1$).

3. Solutions to Trees

As stated in the Introduction, the tree is a very important extension of the star graph and line graph and represents one of the two classes of topologies. It is known that all Markov chains on trees are reversible [21]. Thus, SIMC on trees will be

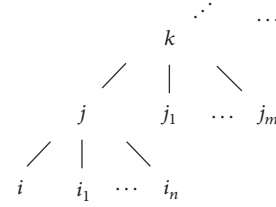


FIGURE 1: Schematic plot of a subchain on a tree. i, i_1, \dots, i_n are leaves and all children of state j ; j, j_1, \dots, j_m are all children of state k .

exploited in this section, although some of them are still not applied in real systems. In this letter, the term “tree” refers to a connected tree. For a Markov chain on a tree with vertices v_1, v_2, \dots, v_m and with a TRM $Q = (q_{ij})$, its state space is composed of the vertices $\{v_1, \dots, v_m\}$, an edge between vertex v_i and vertex v_j implies the nonzero transition rates q_{ij} and q_{ji} , and q_{ij}/q_i is the probability of a transition from v_i to v_j .

It is concluded that all Markov chains on trees are identified by our approach. There are various kinds of trees, so we cannot provide individual proofs and algorithms. Without loss of generality, a complete binary tree is provided to show a general proof and solution. According to the classification of Wolfram MathWorld, there are many valid and specific trees: Banana Tree, Cross Graph, E Graph, Fork Graph, Firecracker Graph, H Graph, Spider Graph, and so forth. Cross Graph, E Graph, Fork Graph, H Graph, and Spider Graph can be conformationally viewed as a Star-Branch Graph (a special tree) which can be identified by the algorithms in earlier publications [19, 20]. Double-Star Graph, Banana Tree, and Firecracker Graph are three particular and regular models of them. For demonstration on solutions of specific trees, those three trees will also be addressed as representatives of trees.

It is mentioned that, for SIMC on trees, the greatest contribution should be to discover the fundamental way and the corresponding algorithm to identify the TRM due to the challenge of performing the reverse operation as stated earlier.

3.1. General Conclusions on the Tree. At the start of this subsection, we present several useful conclusions from the derivative constraints which are convenient to determine the transition rates of Markov chain on trees as the first step.

Lemma 7. *For a subchain on a tree such as in Figure 1, the following conclusions hold.*

(i) One can obtain the quantities $\pi_i, \pi_j, q_i = q_{ij}, q_j, q_{ji}$ at least from the sojourn and hitting time PDFs at state i ; in other words, $\pi_i, \pi_j, q_i = q_{ij}, q_j, q_{ji}$ at least can be expressed in terms of real functions of $E\sigma, c_1, c_2$, and c_3 .

(ii) Further, one can obtain $\pi_i, \pi_j, q_i = q_{ij}, q_j, q_{ji}$ and q_k, q_{jk}, q_{kj} , and π_k at least from the sojourn and hitting time PDFs at state i , provided that the transitions between j and its other children i_1, \dots, i_n , that is, $q_{s,j}$ and $q_{j,s}$ for any $s \in \{i_1, \dots, i_n\}$, are all known or have been identified. That is to say, $\pi_i, \pi_j, q_i = q_{ij}, q_j, q_{ji}$ and q_k, q_{jk}, q_{kj} , and π_k at least can be expressed in terms of real functions of $E\sigma, c_1, \dots, c_5$ and $q_{s,j}, q_{j,s}$ ($s \in \{i_1, \dots, i_n\}$).

Proof. (i) Firstly, $\pi_i, q_i = q_{ij}$ are easy to get by Lemma 6. Secondly, by Corollary 5,

$$\begin{aligned} c_1 &= \pi_j q_{ji}, \\ c_2 &= \pi_j q_{ji}^2, \\ c_3 &= \pi_j q_{ji}^2 q_j. \end{aligned} \quad (11)$$

Thus,

$$\begin{aligned} q_{ji} &= \frac{c_2}{c_1}, \\ \pi_j &= \frac{c_1}{q_{ji}} = \frac{c_1^2}{c_2}, \\ q_j &= \frac{c_3}{c_2}, \end{aligned} \quad (12)$$

which imply the first assertion.

(ii) On the basis of (i), one needs only to find out the algorithm to obtain q_k, q_{jk}, q_{kj} , and π_k . For ease of exposition, set the constants as $X = \sum_{s=i_1}^{i_n} q_{js}$, $Y = \sum_{s=i_1}^{i_n} q_{js} q_{sj}$, and $Z = \sum_{s=i_1}^{i_n} q_{js} q_s q_{sj}$ since $q_s = q_{s,j}$ and $q_{j,s}$ are all known for any $s \in \{i_1, \dots, i_n\}$.

Firstly, one has

$$\begin{aligned} q_{jk} &= q_j - q_{ji} - \sum_{s=i_1}^{i_n} q_{js} = \frac{c_3}{c_2} - \frac{c_2}{c_1} - \sum_{s=i_1}^{i_n} q_{js} \\ &= \frac{c_3}{c_2} - \frac{c_2}{c_1} - X. \end{aligned} \quad (13)$$

Secondly, by Corollary 5,

$$\begin{aligned} c_4 &= \pi_j q_{ji}^2 \left[q_j^2 + q_{jk} q_{kj} + \sum_{s=i_1}^{i_n} q_{js} q_{sj} \right], \\ c_5 &= \pi_j q_{ji}^2 \left[q_j^3 + q_{jk} q_{kj} q_j + 2q_j q_{jk} q_{kj} \right. \\ &\quad \left. + \sum_{s=i_1}^{i_n} (2q_j q_{js} q_{sj} + q_{js} q_s q_{sj}) \right]. \end{aligned} \quad (14)$$

It follows that

$$\begin{aligned} q_{kj} &= \frac{c_4/c_2 - q_j^2 - \sum_{s=i_1}^{i_n} q_{js} q_{sj}}{q_{jk}} \\ &= \frac{c_4/c_2 - c_3^2/c_2^2 - \sum_{s=i_1}^{i_n} q_{js} q_{sj}}{c_3/c_2 - c_2/c_1 - \sum_{s=i_1}^{i_n} q_{js}} \\ &= \frac{c_4/c_2 - c_3^2/c_2^2 - \sum_{s=i_1}^{i_n} q_{js} q_{sj}}{c_3/c_2 - c_2/c_1 - X}, \end{aligned}$$

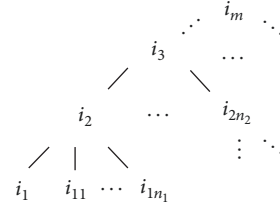


FIGURE 2: Schematic plot of a subtree. $i, i_{11}, \dots, i_{1n_1}$ are leaves and all children of state i_2 ; $i_2, i_{21}, \dots, i_{2n_2}$ are all children of state i_3 ; i_m is the root.

$$\begin{aligned} q_k &= \frac{c_5/c_2 - q_j^3 - \sum_{s=i_1}^{i_n} (2q_j q_{js} q_{sj} + q_{js} q_s q_{sj})}{q_{jk} q_{kj}} - 2q_j \\ &= \frac{c_5/c_2 - c_3^3/c_2^3 - \sum_{s=i_1}^{i_n} (2(c_3/c_2) q_{js} q_{sj} + q_{js} q_s q_{sj})}{c_4/c_2 - \sum_{s=i_1}^{i_n} q_{js} q_{sj} - c_3^2/c_2^2} \\ &\quad - 2\frac{c_3}{c_2} \\ &= \frac{c_5/c_2 - c_3^3/c_2^3 - (2(c_3/c_2)Y + Z)}{c_4/c_2 - c_3^2/c_2^2 - Y} - 2\frac{c_3}{c_2}. \end{aligned} \quad (15)$$

Finally, by reversibility, it easily follows that

$$\begin{aligned} \pi_k &= \frac{\pi_j q_{jk}}{q_{kj}} = \frac{(c_1^2/c_2)(c_3/c_2 - c_2/c_1 - \sum_{s=i_1}^{i_n} q_{js})^2}{c_4/c_2 - c_3^2/c_2^2 - \sum_{s=i_1}^{i_n} q_{js} q_{sj}} \\ &= \frac{(c_1^2/c_2)(c_3/c_2 - c_2/c_1 - X)^2}{c_4/c_2 - c_3^2/c_2^2 - Y}, \end{aligned} \quad (16)$$

which completes the proof. \square

Lemma 8. A subtree such as that in Figure 2 is considered, in which i_m is the root, i_1, \dots, i_{m-1} are the leftmost children, and $i_1, i_{11}, \dots, i_{1n_1}$ are leaves as all children of state i_2 . Let D_k denote the descendants of state i_k for $k = 1, \dots, m-1$. For example, D_1 is empty and $D_2 = \{i_{11}, \dots, i_{1n_1}\}$.

One can obtain $\pi_s, \pi_t, q_s, q_{st}, q_{ts}$ ($s, t = i_1, i_2, \dots, i_m$) from the sojourn and hitting time PDFs at state i_1 , provided that the transitions between i_k and its descendants in D_k for any $k (= 2, 3, \dots, m-1)$ are all known or have been identified.

Proof. Firstly, one can obtain $\pi_{i_1}, \pi_{i_2}, \pi_{i_3}, q_{i_1} = q_{i_1, i_2}, q_{i_2}, q_{i_3}, q_{i_2, i_1}, q_{i_2, i_3}, q_{i_3, i_2}$ according to (ii) of Lemma 7.

Secondly, $q_{i_3, i_4} = q_{i_3} - q_{i_3, i_2} - \sum_{s=1}^{n_2} q_{i_3, i_s}$, and by Corollary 5,

$$\begin{aligned} c_6 &= \pi_{i_2} q_{i_2, i_1}^2 \left[q_{i_2, i_3} q_{i_3, i_4} q_{i_4, i_3} q_{i_3, i_2} \right. \\ &\quad \left. + \sum_{a_1, a_2 \in D_3 \cup \{i_3\} \setminus \{i_1\}} q_{i_2, a_1} q_{a_1, a_2} q_{a_2, a_1} q_{a_1, i_2} \right], \end{aligned} \quad (17)$$

$$c_7 = \pi_{i_2} q_{i_2, i_1}^2 \left[q_{i_2, i_3} q_{i_3, i_4} (\mathbf{q}_{i_4} + q_{i_3} + q_{i_2}) q_{i_4, i_3} q_{i_3, i_2} \right. \\ \left. + \sum_{a_1, a_2, a_3 \in D_3 \cup \{i_3\} \setminus \{i_1\}} q_{i_2, a_1} q_{a_1, a_2} q_{a_3} q_{a_2, a_1} q_{a_1, i_2} \right]. \quad (18)$$

Obviously, only the bold part is unknown in (17). Thus, one can determine q_{i_4, i_3} by (17). Then, only the bold part is unknown in (18), and q_{i_4} can be determined by (18). Furthermore, $\pi_{i_4} = \pi_{i_3} q_{i_3, i_4} / q_{i_4, i_3}$.

Finally, by induction, it is not difficult to verify that $\pi_s, q_s, q_{st}, q_{ts}$ ($s, t = i_5, i_6, \dots, i_m$) can be identified by the subsequent $c_8, c_9, \dots, c_{2m-2}, c_{2m-1}$.

The proof is completed. \square

Remark 9. The above conclusion is true by observation at i_s for any $s = i_1, i_2, \dots, i_{n_1}$ provided that the transitions between i_2 and its other descendants are all known or have been identified.

Lemmas 7 and 8 show that if a state j in a tree has n children and the transitions between states j and $n-1$ of the n children as well as the descendants of such $n-1$ children are all known or have been identified, then the transitions between state j , the n th child (say i), and the parent can be identified by the PDFs of sojourn and hitting time at the n th child (i.e., state i).

3.2. Main Results on Trees. Recall the statement in the Introduction; it is trivial to identify the TRM of an underlying Markov chain provided that every state of this chain is observed. People hope that the number of states observed is as few as possible. But there is no strict criterion for determining how many states are observed to use for a general model other than simplicity and ease of solution. However, there is no doubt that it is also of questionable significance to finish it if the number is more than half of the whole chain. Thus, we impose a mild restriction on the number of observed states such that it is smaller than or equal to half of the whole chain.

First of all, a solution is first provided to complete the binary tree, in which all interior nodes have two children and all leaves have the same depth or same level.

Theorem 10. *If the initial distribution of a Markov chain on a complete binary tree is the equilibrium distribution, then this tree can be identified by the PDFs of the sojourn time and hitting time for all leaves.*

Proof. Consider a complete binary tree with a depth of h ; it has 2^h leaves at the last level h and has $2^h - 1$ nonleaf nodes (including one root).

Now, let us use mathematical induction to prove these facts as follows.

Step 1. We prove the case with 2 levels. For the convenience of expression, let 1–4 be the leaf from the left to the right at the last level, 5–6 be the node from the left to the right

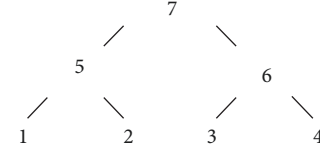


FIGURE 3: Schematic plot of a complete binary tree with two levels. 1 and 2 are leaves and two children of state 5; 3 and 4 are leaves and two children of state 6; state 7 is the root.

at the former level, and 7 be the root (see Figure 3). Let π_i be the corresponding equilibrium distribution for state i for $i = 1, 2, \dots$. Firstly, according to Lemma 6, one can obtain π_1, \dots, π_4 , and $q_1 = q_{15}, q_2 = q_{25}, q_3 = q_{36}$, and $q_4 = q_{46}$ by observation at states 1–4. Secondly, (i) of Lemma 7 implies that $q_5, q_6, q_{51}, q_{52}, q_{63}$, and q_{64} can be solved. Thus, one has $q_{57} = q_5 - q_{51} - q_{52}$ and $q_{67} = q_6 - q_{63} - q_{64}$.

By observation at state 1 (resp., state 3), according to Corollary 5, one has $c_4 = \pi_5(q_{51})^2[q_5^2 + q_{52}q_{25} + q_{57}q_{75}]$ (resp., $c_4 = \pi_6(q_{63})^2[q_6^2 + q_{64}q_{46} + q_{67}q_{76}]$) and $c_5 = \pi_5(q_{51})^2[q_{52}q_{25} + q_{52}q_{25}q_5 + q_{57}q_{75}q_5 + q_{57}q_{77}q_{75}]$. Finally, one gets q_{75} (resp. q_{76}) and q_7 by (ii) of Lemma 7. Those complete the proof of 2 levels.

Step 2. We suppose that this claim is true for any complete binary tree with k (≥ 2) levels. For a complete binary tree with $k+1$ levels, we then show it is correct as follows. At this point, a complete binary tree with k levels can be viewed as a left or right child. This means that all transition rates in two complete binary trees with k levels can be identified by observation at all leaves between 1 and 2^{h+1} .

For conciseness of notation, let L_k (resp. L_{k-1}) be a subtree belonging to the left child in the complete binary tree with $k+1$ (resp., k) levels, and let $0-(k+1)$ be the number of the leftmost nodes at each level from the bottom to the top in such complete binary tree with $k+1$ levels.

Firstly, by observation at state 0, according to Corollary 5, one has

$$c_{2k+1} = \pi_1 (q_{10})^2 \left[\sum_{a_1, \dots, a_{k-1} \neq 0 \in L_k} q_{1a_1} q_{a_1 a_2} \cdots q_{a_{k-2} a_{k-1}} q_{a_{k-1} a_{k-2}} \cdots q_{a_2 a_1} q_{a_1 1} \right] = \pi_1 (q_{10})^2 \\ \cdot \left[\sum_{a_1, \dots, a_{k-1} \neq 0 \in L_{k-1}} q_{1a_1} q_{a_1 a_2} \cdots q_{a_{k-2} a_{k-1}} q_{a_{k-1} a_{k-2}} \cdots q_{a_2 a_1} q_{a_1 1} \right] + \pi_1 (q_{10})^2 [q_{12} q_{23} \cdots q_{k-1, k} q_k q_{k-1, k-2} \cdots q_{32} q_{21}]. \quad (19)$$

By inductive assumptions, there is only one unknown q_k in (19). Thus, q_k can be determined. Again by inductive

assumptions, one can obtain $q_{k,k+1} = q_k - q_{k,k-1} - q_{k,s}$ (where s denotes the right child of the node k).

Analogous to the above, one can get $q_{k+1,k}$ by

$$c_{2k+2} = \pi_1 (q_{10})^2 \left[\sum_{a_1, \dots, a_k \neq 0 \in L_k} q_{1a_1} q_{a_1 a_2} \cdots q_{a_{k-1} a_k} q_{a_k a_{k-1}} \right. \\ \left. \cdots q_{a_2 a_1} q_{a_1 1} \right] = \pi_1 (q_{10})^2 \left[\sum_{a_1, \dots, a_k \neq 0 \in L_{k-1}} q_{1a_1} q_{a_1 a_2} \right. \\ \left. \cdots q_{a_{k-1} a_k} q_{a_k a_{k-1}} \right] + \pi_1 (q_{10})^2 [q_{12} q_{23} \cdots q_{k,k+1} q_{k+1,k} \\ \cdots q_{32} q_{21}]. \quad (20)$$

Thus, $q_{k+1,k}$ can be determined by (20).

Those imply that all transition rates from the left subtree (for the tree with $k+1$ levels) can be identified by observations at the first 2^{k-1} leaves.

Secondly, one can identify the rest (i.e., right part) of the tree with $k+1$ levels by duplicate method of the left part.

These have proved that the conclusions of induction are true for a complete binary tree with $k+1$ levels.

As a matter of fact, this theorem can also be proved simply by Lemma 8. \square

Remark 11. This proof shows the corresponding algorithm to determine all transition rates. The solution by observation at all leaves is satisfied by the mild restriction on the number of observed states, although the number of leaves is bigger than the number of nonleaf nodes by one. Because it also can be identified by observation at any $2^h - 1$ of those 2^h leaves, for example, without loss of generality, assume that 1–3 are observable for a tree in Figure 3; one can obtain $q_1 = q_{15}, q_{51}, q_2 = q_{25}, q_{52}, q_5, q_{57}, q_7, q_{75}, \pi_1, \pi_2, \pi_5, \pi_7$ by observations at 1 and 2 and $q_3 = q_{36}, q_{63}, q_6, \pi_3, \pi_6$ by observations at 3 and then obtain q_{76}, q_{67}, q_6 by observation

at 1 or 2; thus, $q_{64} = q_6 - q_{67} - q_{63}, \pi_4 = 1 - \pi_1 - \pi_2 - \pi_3 - \pi_5 - \pi_6 - \pi_7$, and $q_{46} = \pi_6 q_{64} / \pi_4$. If the number of all leaves is less than that of all nonleaf nodes for a tree, such as a tree in which many nodes have only one child, then this solution is very valid. These assertions are valid to n -tree later.

One can carry this line of argument a bit further to obtain the following result.

Theorem 12. *If the initial distribution of a Markov chain on n -tree is the equilibrium distribution, then this tree can be identified by the PDFs of the sojourn time and hitting time for all leaves.*

One may proceed as previously explained to obtain a more general conclusion for the tree.

Theorem 13. *If the initial distribution of a Markov chain on a tree is the equilibrium distribution, then this tree can be identified by the PDFs of the sojourn time and hitting time for all leaves.*

Theoretically, the transition rates about state i in a tree can be calculated by PDFs at any leaf from the descendants. However, the one from the shortest path should be optimal to minimize the error propagation.

Furthermore, identification by observation at leaves is just one of the solutions for Markov chains on trees. There are also other solutions for particular trees; for example, the number of leaves is bigger than the number of nonleaf nodes by one.

3.3. Solutions to Special Trees. In this subsection, Double-Star Graph, Banana Tree, and Firecracker Tree, as representatives of trees, will be addressed for demonstration to solutions of specific trees.

3.3.1. Double-Star Graph. Consider a Double-Star Graph Markov chain $\{X_t; t \geq 0\}$ (see Figure 4), where the two central states O_1 and O_2 are observable states. The transition rate matrix is given as follows:

$$Q = \begin{pmatrix} q_0 & q_{01} & q_{02} & \cdots & q_{0,M-1} & q_{0,M} & 0 & 0 & \cdots & 0 \\ q_{10} & q_1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ q_{20} & 0 & q_2 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{M-1,0} & 0 & 0 & \cdots & q_{M-1} & 0 & 0 & 0 & \cdots & 0 \\ q_{M,0} & 0 & 0 & \cdots & 0 & q_M & q_{M,M+1} & q_{M,M+2} & \cdots & q_{M,N} \\ 0 & 0 & 0 & \cdots & 0 & q_{M+1,M} & q_{M+1} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & q_{M+2,M} & 0 & q_{M+2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & q_{N,M} & 0 & 0 & \cdots & q_N \end{pmatrix}. \quad (21)$$

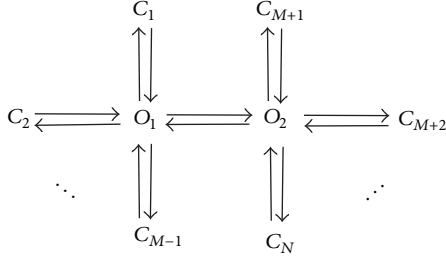


FIGURE 4: Schematic plot of Markov chain on a Double-Star Graph.

One can observe the two central states $0 = O_1 \in \mathcal{O}_1$ and $M = O_2 \in \mathcal{O}_2$.

For the conciseness of notation, we denote $i = C_i$ ($i = 1, \dots, M-1$), $j = C_j$ ($j = M+1, \dots, N$), and $0 = O_1, M = O_2$. Thus, the state space is simple for $\mathbf{S} = \{0, 1, \dots, M-1, M, M+1, \dots, N\}$ and the transition rates q_{ij} are nonzero only for $i \in 0, M$ or $j \in 0, M$. Thus, one intends to derive q_{i0}, q_{0i} for all $i = 1, \dots, M-1$ and q_{jM}, q_{Mj} for all $j = M+1, \dots, N$, as well as q_{0M}, q_{M0} . We set

$$\begin{aligned} \pi_0 &= \left(1 + \frac{q_{0M}}{q_{M0}} + \sum_{i=1}^{M-1} \frac{q_{0i}}{q_{i0}} + \sum_{j=M+1}^N \frac{q_{Mj}}{q_{jM}} \right)^{-1}, \\ \pi_i &= \frac{\pi_0 q_{0i}}{q_{i0}}, \quad 1 \leq i \leq M-1; \\ \pi_j &= \frac{\pi_0 q_{0M} q_{Mj}}{q_{jM} q_{M0}}, \quad M+1 \leq j \leq N; \\ \pi_M &= \frac{\pi_0 q_{0M}}{q_{M0}}. \end{aligned} \quad (22)$$

Then, $\{\pi_0, \dots, \pi_{M-1}, \pi_M, \pi_{M+1}, \dots, \pi_N\}$ is the equilibrium distribution and the reversibility in (1) is satisfied.

From now on, we investigate the algorithm of how to identify such chain.

Firstly, we observe state 0 and let $\sigma^{(0)}$ be its sojourn time. It follows from Corollary 2 that

$$\begin{aligned} q_0 &= \frac{1}{E\sigma^{(0)}}, \\ \pi_0 &= c_1^{(0)} E\sigma^{(0)}. \end{aligned} \quad (23)$$

Secondly, we observe the state M and let $\sigma^{(M)}$ be its sojourn time. It follows from Corollary 2 that

$$\begin{aligned} q_M &= \frac{1}{E\sigma^{(M)}}, \\ \pi_M &= c_1^{(M)} E\sigma^{(M)}. \end{aligned} \quad (24)$$

Finally, we observe the two states 0 and M , that is, $\mathbf{O} = \{0, M\}$ ($\mathbf{C} = \mathbf{S} - \mathbf{O}$). Let $\tau^{(0,M)}$ (resp., $\sigma^{(0,M)}$) be the hitting time (resp., sojourn time) of \mathbf{O} . According to Theorem 1, one can get the following lemma.

Lemma 14. The PDF of the hitting time $\tau^{(0,M)}$ is, for $t > 0$, as follows:

$$f_\tau(t) = \sum_{i=1}^{M-1} \gamma_i \exp(-\alpha_i t) + \sum_{j=M+1}^N \gamma_j \exp(-\alpha_j t), \quad (25)$$

where

$$\begin{aligned} \alpha_i &= q_{i0} = q_i, \quad 1 \leq i \leq M-1; \\ \alpha_j &= q_{jM} = q_j, \quad M+1 \leq j \leq N; \\ \gamma_i &= \frac{\pi_0}{1 - (\pi_0 + \pi_M)} q_{0i}, \quad 1 \leq i \leq M-1; \\ \gamma_j &= \frac{\pi_M}{1 - (\pi_0 + \pi_M)} q_{Mj}, \quad M+1 \leq j \leq N. \end{aligned} \quad (26)$$

Proof. Here, $\mathbf{C} = \{1, \dots, M-1, M+1, \dots, N\}$. So,

$$Q_{cc} = \begin{pmatrix} q_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & q_2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{M-1} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & q_{M+1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & q_{M+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & q_N \end{pmatrix}. \quad (27)$$

According to the proof of Theorem 1, $-\alpha_i = -q_i$ ($1 \leq i \leq M \leq N$) are all real eigenvalues of Q_{cc} , and

$$\begin{aligned} P(\tau^{(0,M)} > t) &= \sum_{i \neq 0, M} \frac{\pi_i}{1 - (\pi_0 + \pi_M)} \sum_{j \neq 0, M} p_{ij}^c(t) \\ &= \frac{1}{1 - (\pi_0 + \pi_M)} \left[\sum_{i=1}^{M-1} \pi_i \exp(-\alpha_i t) \right. \\ &\quad \left. + \sum_{j=M+1}^N \pi_j \exp(-\alpha_j t) \right]. \end{aligned} \quad (28)$$

Performing the differential calculation in (28), by reversibility,

$$\begin{aligned} \gamma_i &= \frac{\pi_i \alpha_i}{1 - (\pi_0 + \pi_M)} = \frac{\pi_i q_{i0}}{1 - (\pi_0 + \pi_M)} \\ &= \frac{\pi_0}{1 - (\pi_0 + \pi_M)} q_{0i}, \quad 1 \leq i \leq M-1; \\ \gamma_j &= \frac{\pi_j \alpha_j}{1 - (\pi_0 + \pi_M)} = \frac{\pi_j q_{jM}}{1 - (\pi_0 + \pi_M)} \\ &= \frac{\pi_M}{1 - (\pi_0 + \pi_M)} q_{Mj}, \quad M+1 \leq j \leq N. \end{aligned} \quad (29)$$

This completes the proof. \square

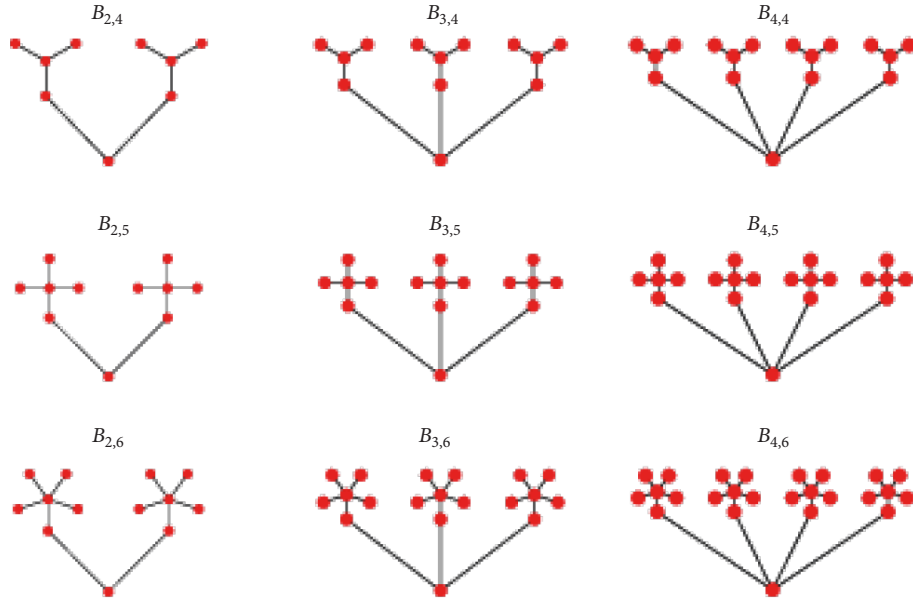


FIGURE 5: Schematic plot of banana trees (adapted from Wolfram MathWorld).

The following theorem is an immediate consequence of Lemma 14.

Theorem 15. *If the initial distribution of the Markov chain with Double-Star Graph in Figure 4 is the equilibrium distribution, then it can be identified by the PDFs of the sojourn time and hitting time for the two central states O_1 and O_2 .*

Proof. Firstly, q_0, π_0 and q_M, π_M can be obtained by (23) and (24). Secondly, according to Lemma 14, q_{i0} ($1 \leq i \leq M-1$) and q_{jM} ($M+1 \leq j \leq N$) can be obtained by the first two assertions of (26), because all α_i and γ_i are known from the PDF in Lemma 14 by observation of states 0 and M . Finally, q_{0i} ($1 \leq i \leq M-1$) and q_{Mj} ($M+1 \leq j \leq N$) can be followed by the final two assertions of (26). \square

Note that this proof shows the corresponding algorithm to determine all transition rates of such Double-Star Graph chain. In the course of calculation, one can easily derive real roots from comparatively accurate PDFs.

3.3.2. Banana Tree Graph. An (n, k) -banana tree, say $B_{n,k}$, as defined by Chen et al. [23], is a graph obtained by connecting one leaf of each of n copies of a k -star graph with a single root vertex that is distinct from all the stars. Thus, there are $n(k-2)$ leaves, $2n$ nodes, and 1 root. This is one of the regular trees.

Consider a Markov chain $\{X_t; t \geq 0\}$ with Banana Tree Graph $B_{n,k}$ (see Figure 5).

For $k = 2$ or $k = 3$, it is degenerated into the special case of Star-Branch Graph.

For $k = 4$, there are two leaves for each k -star graph. It is implied that the number $2n$ of all leaves is less than 1, compared with $2n+1$ of all nonleaf nodes. So, observation at all leaves is valid. According to Theorem 13, a normal solution is as follows.

Theorem 16. *If the initial distribution of a Markov chain with Banana Tree Graph $B_{n,k}$ ($k \leq 4$) is the equilibrium distribution, then it can be identified by the PDFs of the sojourn time and hitting time for all leaves.*

For $k \geq 5$, the number of all leaves is bigger than that of nonleaf nodes. An optional solution is as follows.

Theorem 17. *If the initial distribution of a Markov chain with Banana Tree Graph $B_{n,k}$ ($k \geq 5$) is the equilibrium distribution, then it can be identified by the PDFs of the sojourn time and hitting time for the root state and all node states, that is, all nonleaf states.*

Proof. For ease of exposition, suppose that r is the root (see the previous introduction of banana tree in this subsection) of this tree and that o is the central state, l is the leaf, and s are the others at each k -star graph, respectively. Firstly, the corresponding $q_r, \pi_r, q_o, \pi_o, q_s,$ and π_s can be obtained by each nonleaf state. Secondly, according to the proof of Double-Star Graph, all $q_l = q_{lo}, q_{ol}$ can be obtained by observation at all nonleaf states, because their Q_{cc} is similar to that of observation at $\{O_1, O_2\}$ for Double-Star Graph. Finally, for each k -star graph, $q_{os} = q_o - \sum_l q_{ol}, q_{so} = \pi_o q_{os} / \pi_s, q_{sr} = q_s - q_{so},$ and $q_{rs} = \pi_s q_{sr} / \pi_r.$ \square

3.3.3. Firecracker Graph. An (n, k) - firecracker, say $F_{n,k}$, is a graph obtained by the concatenation of n k -stars by linking one leaf from each [23].

A Markov chain $\{X_t; t \geq 0\}$ with Firecracker Graph $F_{n,k}$ (see Figure 6) is provided.

For $k = 2$ or $k = 3$, it is degenerated into the special tree which is similar, but not completely equivalent, to a Star-Branch Graph.

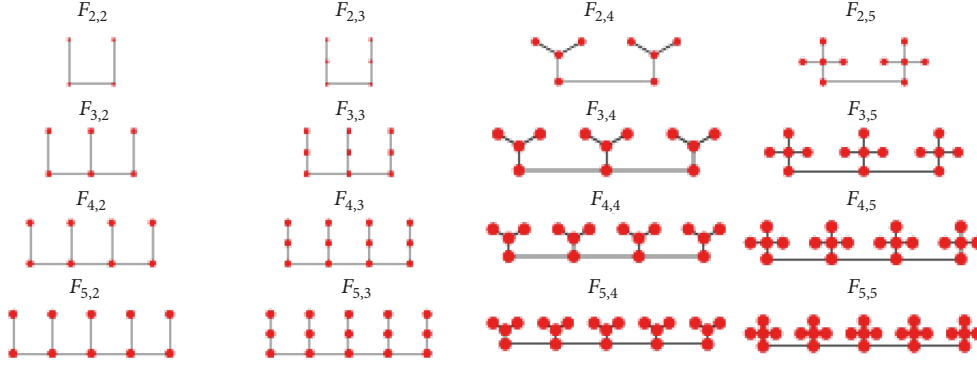


FIGURE 6: Schematic plot of firecracker graphs (adapted from Wolfram MathWorld).

For $k \leq 4$, there are two leaves for each k -star graph. It is implied that the number $2n$ of all leaves is identical to that of all nonleaf nodes. So, observation at all leaves is valid. According to Theorem 13, a normal solution for $k \leq 4$ is as follows.

Theorem 18. *If the initial distribution of a Markov chain with Firecracker Graph $F_{n,k}$ ($k \leq 4$) is the equilibrium distribution, then it can be identified by the PDFs of the sojourn time and hitting time for all leaves.*

For $k \geq 5$, the number of all leaves is bigger than that of nonleaf nodes. Following the technique of proof in Theorem 17, it is easy to verify such optional solution.

Theorem 19. *If the initial distribution of a Markov chain with Firecracker Graph $F_{n,k}$ ($k \geq 5$) is the equilibrium distribution, then it can be identified by the PDFs of the sojourn time and hitting time for all nonleaf states.*

3.4. Numerical Example. To demonstrate how to apply our algorithms to real data, we present a numerical example here. As we mentioned before, we focus on the final step and then we omit the preprocessing of data. Thus, the data we have is the observed PDFs of sojourn time and hitting time at each leaf. Based upon these data, we could find all the transition rates of this tree.

Example 1. The model of a Markov chain is a binary tree in Figure 3 with the true transition rates matrix $Q = (q_{ij})_{S \times S}$ given by

$$Q = \begin{pmatrix} -10 & 0 & 0 & 0 & 10 & 0 & 0 \\ 0 & -25 & 0 & 0 & 25 & 0 & 0 \\ 0 & 0 & -30 & 0 & 0 & 30 & 0 \\ 0 & 0 & 0 & -15 & 0 & 15 & 0 \\ 5 & 15 & 0 & 0 & -50 & 0 & 30 \\ 0 & 0 & 30 & 50 & 0 & -100 & 20 \\ 0 & 0 & 0 & 0 & 90 & 60 & -150 \end{pmatrix}. \quad (30)$$

We can divide the calculation into two steps: fitting sojourn time and hitting time histogram and transition rates estimation.

Step 1 (sojourn time and hitting time histogram estimate). Suppose that we have obtained the PDFs by fitting (simulation) as follows:

$$\begin{aligned} f_{\tau^{(1)}}(t) = & 0.002032e^{-186.015025t} + 0.002801e^{-107.093861t} \\ & + 0.031748e^{-43.239654t} \\ & + 0.530183e^{-0.537764t} \\ & + 0.121303e^{-9.103660t} \\ & + 0.000006e^{-24.010036t} \quad (t \geq 0), \end{aligned} \quad (31)$$

$$\begin{aligned} f_{\tau^{(2)}}(t) = & 0.018133e^{-185.792813t} + 0.023965e^{-106.751426t} \\ & + 1.024495e^{-1.089438t} \\ & + 0.265559e^{-8.032879t} \\ & + 0.669767e^{-30.000000t} \\ & + 0.091104e^{-23.333445t} \quad (t \geq 0), \end{aligned} \quad (32)$$

$$\begin{aligned} f_{\tau^{(3)}}(t) = & 0.098146e^{-185.213916t} + 1.004750e^{-98.503674t} \\ & + 0.063878e^{-43.082231t} \\ & + 2.304378e^{-2.701274t} \\ & + 0.960962e^{-7.138694t} \\ & + 0.001383e^{-13.360211t} \quad (t \geq 0), \end{aligned} \quad (33)$$

$$\begin{aligned} f_{\tau^{(4)}}(t) = & 0.425895e^{-185.406369t} + 3.907008e^{-101.419933t} \\ & + 0.117730e^{-43.639996t} \\ & + 2.473349e^{-3.548045t} \end{aligned}$$

$$\begin{aligned}
& + 3.456848e^{-18.354848t} \\
& + 0.897366e^{-12.630809t} \quad (t \geq 0),
\end{aligned} \tag{34}$$

$$\begin{aligned}
f_{\sigma^{(1)}}(t) &= 25e^{-25t}, \\
f_{\sigma^{(2)}}(t) &= 10e^{-10t}, \\
f_{\sigma^{(3)}}(t) &= 15e^{-50t}, \\
f_{\sigma^{(4)}}(t) &= 30e^{-50t}
\end{aligned} \tag{35}$$

$$(t \geq 0),$$

where $f_{\sigma^{(i)}}(t)$ (resp., $f_{\tau^{(i)}}(t)$) is the PDF of sojourn time (resp., hitting time) at the state i for $i = 1, 2, 3, 4$.

Step 2 (transition rate calculations). Firstly, by (35) and (10), we have

$$\begin{aligned}
q_1 &= q_{15} = 10, \\
q_2 &= q_{25} = 25, \\
q_3 &= q_{36} = 30, \\
q_4 &= q_{46} = 15.
\end{aligned} \tag{36}$$

For state 1, let

$$\begin{aligned}
\alpha_1 &= 186.015025, \\
\alpha_2 &= 107.093861, \\
\alpha_3 &= 43.239654, \\
\alpha_4 &= 0.537764, \\
\alpha_5 &= 9.103660, \\
\alpha_6 &= 24.010036, \\
\gamma_1 &= 0.002032, \\
\gamma_2 &= 0.002801, \\
\gamma_3 &= 0.031748, \\
\gamma_4 &= 0.530183, \\
\gamma_5 &= 0.121303, \\
\gamma_6 &= 0.000006.
\end{aligned} \tag{37}$$

Because $d_1^{(1)} = \sum_{i=1}^6 \gamma_i = 0.688073$, hence $\pi_1 = d_1^{(1)} / (q_1 + d_1^{(1)}) = 0.064378$, $1 - \pi_1 = 0.935622$. Since

$$\begin{aligned}
d_n^{(1)} &= \sum_{i=1}^6 \gamma_i \alpha_i^{n-1}, \\
c_n^{(1)} &= (1 - \pi_1) d_n^{(1)} = 0.935622 d_n,
\end{aligned} \tag{38}$$

we have

$$\begin{aligned}
c_1^{(1)} &= 0.643777, \\
c_2^{(1)} &= 3.218884, \\
c_3^{(1)} &= 160.944206, \\
c_4^{(1)} &= 17945.278970.
\end{aligned} \tag{39}$$

Similarly, we can get for states 2, 3, and 4, respectively,

$$\begin{aligned}
\pi_2 &= 0.077253, \\
\pi_3 &= 0.128755, \\
\pi_4 &= 0.429185, \\
c_1^{(2)} &= 1.931330, \\
c_2^{(2)} &= 28.969957, \\
c_3^{(2)} &= 1448.497854, \\
c_4^{(2)} &= 152092.274678, \\
c_1^{(3)} &= 3.862661, \\
c_2^{(3)} &= 115.879828, \\
c_3^{(3)} &= 11587.982833, \\
c_4^{(3)} &= 1384763.948498, \\
c_1^{(4)} &= 6.437768, \\
c_2^{(4)} &= 321.888412, \\
c_3^{(4)} &= 32188.841202, \\
c_4^{(4)} &= 3894849.785408.
\end{aligned} \tag{40}$$

Thus, we can obtain by the algorithm (proof) of Theorem 10 that

$$\begin{aligned}
q_{51} &= \frac{c_2^{(1)}}{c_1^{(1)}} = 5.000000, \\
q_{63} &= \frac{c_2^{(3)}}{c_1^{(3)}} = 29.999999, \\
q_{52} &= \frac{c_2^{(2)}}{c_1^{(2)}} = 15.000000, \\
q_{64} &= \frac{c_2^{(4)}}{c_1^{(4)}} = 50.000000, \\
q_5 &= \frac{c_3^{(1)}}{c_2^{(1)}} = 50.000000,
\end{aligned}$$

$$q_6 = \frac{c_3^{(3)}}{c_2^{(3)}} = 100.000000,$$

$$\pi_5 = \frac{[c_1^{(1)}]^2}{c_2^{(1)}} = 0.128755,$$

$$\pi_6 = \frac{[c_1^{(3)}]^2}{c_2^{(3)}} = 0.128755,$$

$$q_{57} = q_5 - q_{51} - q_{52} = 30.000000,$$

$$q_{67} = q_6 - q_{63} - q_{64} = 20.000000,$$

$$q_{75} = \frac{c_4^{(1)}/c_2^{(1)} - q_{52}q_{25} - q_5^2}{c_3^{(1)}/c_2^{(1)} - c_2^{(1)}/c_1^{(1)} - q_{52}} = 90.000007,$$

$$q_{76} = \frac{c_4^{(3)}/c_2^{(3)} - q_{64}q_{46} - q_6^2}{c_3^{(3)}/c_2^{(3)} - c_2^{(3)}/c_1^{(3)} - q_{64}} = 59.999999,$$

$$\pi_7 = 1 - \sum_{i=1}^6 \pi_i = 0.042919.$$

(41)

Thus, the TRM of $\{X_t : t \geq 0\}$ is identified as follows:

$$Q = \begin{pmatrix} -10 & 0 & 0 & 0 & 10 & 0 & 0 \\ 0 & -25 & 0 & 0 & 25 & 0 & 0 \\ 0 & 0 & -30 & 0 & 0 & 30 & 0 \\ 0 & 0 & 0 & -15 & 0 & 15 & 0 \\ 5 & 15 & 0 & 0 & -50 & 0 & 30 \\ 0 & 0 & 29.999999 & 50 & 0 & -99.999999 & 20 \\ 0 & 0 & 0 & 0 & 90.000007 & 59.999999 & -150.000006 \end{pmatrix}. \quad (42)$$

Comparing matrix (42) with the original Q -matrix (30), it is not difficult to find that our approach is very efficient under the condition that we obtain an accurate PDF of the hitting time and sojourn time at each leaf.

Note that if Q is symmetric, then the corresponding algorithms are also valid. As a matter of fact, the calculation can be more concise due to the identical equilibrium distributions. In addition, it also can be identified by observations at any three states of four leaves as stated in Remark 11 of Theorem 10.

4. Discussion

In the current paper, we have focused on the statistical identification of Markov chains on trees. As stated earlier, all of them are identified only by the derivative constraints. For example, a complete binary tree is identified by observations at all leaves (see Theorem 10). Many of them are from current applications and others are still not (e.g., in ion channels). However, it is necessary for the theoretical point of view to be extended to the more general one-tree model. It is mentioned that the states of each subset observed should be distinguishable on the experimental recording of applications.

In the future, we might shed light on the statistical identification of reversible Markov chains with cyclic graphs, as another important class analogous to trees such as Tadpole Graph, Pan Graph, and even the more general one-Markov chain. Once reversible Markov chains with cyclic graphs are solved, our approach is then feasible to most of the reversible Markov chains, except those with very complicated network architectures. Hence, the work further opens up the

possibility of carrying out the statistical identification for general Markov chains with reversibility.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants nos. 11671132 and 61403136), the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (2015), Provincial Natural Science Foundation of Hunan (Grant no. 2018JJ2010), Philosophy and Social Science Foundation of Hunan Province (Grant no. 16YBA053), and Key Scientific Research Project of Hunan Provincial Education Department (Grant no. 15A032).

References

- [1] D. Colquhoun and A. G. Hawkes, "On the Stochastic Properties of Bursts of Single Ion Channel Openings and of Clusters of Bursts," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 300, no. 1098, pp. 1–59, 1982.
- [2] F. G. Ball, R. K. Milne, and G. F. Yeo, "Marked continuous-time Markov chain modelling of burst behaviour for single ion channels," *Journal of Applied Mathematics and Decision Sciences*, vol. 2007, Article ID 48138, 2007.
- [3] R. J. Elliot, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*, Springer-Verlag New York, Berlin, Germany, 1995.

- [4] C. Shelley and K. L. Magleby, "Linking exponential components to kinetic states in Markov models for single-channel gating," *The Journal of General Physiology*, vol. 132, no. 2, pp. 295–312, 2008.
- [5] R. J. French and W. F. Wonderlin, "Software for acquisition and analysis of ion channel data: Choices, tasks, and strategies," *Methods in Enzymology*, vol. 207, pp. 711–728, 1992.
- [6] N. P. Jewell, "Mixtures of exponential distributions," *The Annals of Statistics*, vol. 10, no. 2, pp. 479–484, 1982.
- [7] L. C. Keatinge, "Modelling losses with the mixed exponential distribution," in *Proceedings of the Proceeding of the Causality Actuarial Society*, vol. 86, pp. 654–698, 1999.
- [8] D. Colquhoun, C. J. Hatton, and A. G. Hawkes, "The quality of maximum likelihood estimates of ion channel rate constants," *The Journal of Physiology*, vol. 547, no. 3, pp. 699–728, 2003.
- [9] R. Horn and K. Lange, "Estimating kinetic constants from single channel data," *Biophysical Journal*, vol. 43, no. 2, pp. 207–223, 1983.
- [10] B. Larget, "A canonical representation for aggregated Markov processes," *Journal of Applied Probability*, vol. 35, no. 2, pp. 313–324, 1998.
- [11] M. Wagner, S. Michalek, and J. Timmer, "Estimating transition rates in aggregated Markov models of ion channel gating with loops and with nearly equal dwell times," *Proceedings of the Royal Society B Biological Science*, vol. 266, no. 1431, pp. 1919–1926, 1999.
- [12] O. Flomenbom and R. J. Silbey, "Utilizing the information content in two-state trajectories," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 29, pp. 10907–10910, 2006.
- [13] O. Flomenbom and R. J. Silbey, "Toolbox for analyzing finite two-state trajectories," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 78, no. 6, Article ID 066105, 2008.
- [14] D. R. Fredkin and J. A. Rice, "On aggregated Markov processes," *Journal of Applied Probability*, vol. 23, no. 1, pp. 208–214, 1986.
- [15] B. Sakmann and E. Neher, *Single-Channel Recording*, Springer, New York, NY, USA, 2nd edition, 2009.
- [16] C. P. Amann, T. Schmiedl, and U. Seifert, "Communications: Can one identify nonequilibrium in a three-state system by analyzing two-state trajectories?" *The Journal of Chemical Physics*, vol. 132, no. 4, Article ID 041102, 2010.
- [17] C. Jia and Y. Chen, "A second perspective on the Amann-Schmiedl-Seifert criterion for non-equilibrium in a three-state system," *Journal of Physics A: Mathematical and Theoretical*, vol. 48, no. 20, Article ID 205001, pp. 1–12, 2015.
- [18] Y. Deng, S. Peng, M. Qian, and J. Feng, "Identifying transition rates of ionic channels via observations at a single state," *Journal of Physics A: Mathematical and General*, vol. 36, no. 5, pp. 1195–1212, 2003.
- [19] X. Xiang, X. Yang, Y. Deng, and J. Feng, "Identifying transition rates of ionic channels of star-graph branch type," *Journal of Physics A: Mathematical and General*, vol. 39, no. 30, pp. 9477–9491, 2006.
- [20] X. Xiang, D. Yingchun, and Y. Xiangqun, "Markov chain inversion approach to identify the transition rates of ion channels," *Acta Mathematica Scientia*, vol. 32, no. 5, pp. 1703–1718, 2012.
- [21] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, New York, NY, USA, 1979.
- [22] Z. K. Wang and X. Q. Yang, *Birth and death processes and Markov chains*, Springer-Verlag, Berlin; Science Press Beijing, Beijing, 1992.
- [23] W.-C. Chen, H.-I. Lu, and Y.-N. Yeh, "Operations of interlaced trees and graceful trees," *Southeast Asian Bulletin of Mathematics*, vol. 21, no. 4, pp. 337–348, 1997.

