Determination of Markov Model Dynamics from Equilibrium Data Snapshots

Austen Mazenko

under the direction of

Dominic Skinner
Massachusetts Institute of Technology
Graduate Student in Mathematics

Research Science Institute
July 30, 2019

Abstract

Often, when using a Markov State Model (MSM) to model a physical or biological system, only the equilibrium distribution is experimentally measurable, yet this equilibrium alone is insufficient to uniquely fix the system's transition probabilities. To determine these probabilities and thus the dynamics of such systems, this paper considers inhibiting various transitions and using the new equilibria to gain information about the system. We completely determine the minimum number of cuts required to fully characterize three-state systems, and conjecture that $n-1$ cuts is both necessary and sufficient for complete, n-state systems. Because such a characterization is inherently valid only up to scaling, we establish the number of blocks in the transition graph as a lower bound on the degrees of freedom. Finally, we simulate systems to confirm the practicality of our minimum-cut algorithm for the three- and four-state situations.

Summary

Many real-world systems, from the stock market to protein folding and gene expression, have some amount of randomness. We can often describe these systems by a Markov State Model (MSM), which assumes the very next state of something in the system, say a cell or particle, depends only on its present state. Such systems always end up in a steady-state, or equilibrium, situation. Because the equilibrium is often all we can measure, it is of interest to determine all we can about such systems solely from data describing the equilibrium. To do this, we block off transitions between different states in the system, which gives a new equilibrium and more information. We determine how many blocked transitions are sufficient to uniquely determine the behavior of the system, and run simulations to test our algorithm in a realistic situation with natural errors in data. Finally, we also prove the impossibility of determining exactly how quickly the system progresses towards the equilibrium if we may only block transitions, and we provide a bound on exactly how many such pieces of information we cannot determine.

1 Introduction

Stochastic processes, or sequences of random variables, form a central tenet of probability theory with various interpretations in physical systems. Brownian motion [1], protein folding [2], gene expression [3, 4], and evolution [5] represent only some of the real-world dynamical systems exhibiting such randomness. A significant subset of those systems can be further described by a Markov state model (MSM), a model which satisfies the *Markov property*: the state of a given variable or particle one time step in the future depends only on the current state. Indeed, the use of such MSMs, which cluster data into states and then consider transition probabilities between these states, has recently increased in popularity [6] especially in modeling protein folding [7–11] and gene expression [5, 12]. Knowledge of transition probabilities uniquely determines the model, enabling understanding of crucial attributes of systems such as mean first passage times and state differentiation probabilities [8, 13].

In practice, calculating transition probabilities is difficult, as collecting relevant data is challenging, especially in microbiological systems. A recent breakthrough has been made in the methods of data collection by Rosenberg et al. [14], who introduce a method of analyzing thousands of cells in a system at once. This allows us to gain accurate knowledge of equilibria without extrapolating from the states of only a few cells. Although these single-cell snapshots are becoming more obtainable, their static nature limits the recoverable information about dynamics or progression over time. A single set of equilibrium distribution data may represent various transition matrices and systems; that is, such matrices are non-uniquely determined by the individual measurements. This paper's goal is to determine which manipulations of the system (e.g. inhibiting select transitions, isolating states) are necessary and sufficient to uniquely determine the transition matrix and characterize the system.

Some information about the systems is inherently unknowable when only equilibrium information is gathered. In particular, transition probabilities are only determinable up to at least one scaling factor, so, for example, the rate of convergence to equilibrium is indeterminable. While statistical approaches such as Bayesian inference and maximum likelihood estimation have previously been effectively used to reconstruct the most likely transition matrix given data [15, 16], the systems to which they have been applied are time resolved. We instead focus on the measurements and changes to the system that are necessary to algebraically determine transition probabilities when only equilibrium data is available. Furthermore, we do not assume detailed balance, meaning the systems we focus on do not necessarily possess time reversibility. While this statistical mechanics assumption of reversibility simplifies analysis and is commonly used [8, 16], many biological systems do not obey thermodynamic equilibrium and thus do not possess reversibility.

The purpose of this paper is to derive the transition probabilities of MSMs only using knowledge of equilibrium distributions. More precisely, we consider cutting, or inhibiting, different transitions between states. Such cuts change the system and induce different equilibrium distributions among the states, and we use the values of these different distributions together to determine the initial, uncut system's transition probabilities. In Section 2, we formalize the specific attributes of the MSMs considered in this paper and introduce the idea of a cut. In Section 3.1, we focus on complete systems under the assumption of overdamping, so that when a transition is cut its probability shifts to the transition back into the same state. Establishing optimal measurements in these complete systems, in Section 3.2 we characterize all of the three-state degenerate situations, or those having certain inter-state transitions with zero probability. Then, in Section 4, we clarify the issue of scalability by establishing a lower bound on the number of indeterminable scaling factors of transition probabilities in any given system. In Section 5, we simulate three- and four-state complete systems to determine the feasibility of our explicit algorithm on noisy data. Finally, we defer proofs of some of our results to the appendices for space purposes.

2 Formalization of Markov Processes

Let $\mathbb{P}(A|B)$ denote the conditional probability of event A happening given B. Formally, a process is *Markovian* given the following: whenever X_i are random variables in, and s_i are elements of, a state space S, we have

$$\mathbb{P}(X_{n+1} = s_{n+1}|X_n = s_n, X_{n-1} = s_{n-1}, ..., X_0 = s_0) = \mathbb{P}(X_{n+1} = s_{n+1}|X_n = s_n).$$

Focusing only on finite state spaces S, we can thus define the n-step transition probability $p_{i,j}(n) := \mathbb{P}(X_n = j|X_0 = i)$ for $i, j \in S$. This gives rise to the transition matrix

$$P = \begin{bmatrix} p_{1,1}(1) & \dots & p_{1,m}(1) \\ \vdots & \ddots & \vdots \\ p_{m,1}(1) & \dots & p_{m,m}(1) \end{bmatrix},$$

where $|S| = m$. Note that $\sum\limits_{k\in S} p_{i,k}(1) = \sum\limits_{k\in S} \mathbb{P}(X_{n+1} = k|X_n = i) = \mathbb{P}\left(\cup_{k\in S} X_{n+1} = k|X_n = i\right) = 1$, which is to say P is *stochastic*, meaning each of its row sums is 1.

In a given Markov process with transition matrix P, we consider an initial distribution $\pi \in \mathbb{R}^m$ over the different states, where the ith element of π represents the probability of starting at state i. In particular, if the system is one of cells and the states are different states the cells can be in, the ith element of π denotes the proportion of all the cells which will start in state i. Thus, after one time step, the expected distribution of the cells is πP. Now, the Markov property also gives $p_{i,j}(n)$ in terms of shorter time step lengths as demonstrated by the Chapman-Kolmogorov Equation, the validity of which is shown, e.g., by Meyn and Tweedie [17].

Lemma 2.1 (Chapman-Kolmogorov)**.** *For a Markov process with n-step transition probability $p_{i,j}(n)$ and integers $n_1, n_2 > 0$, the equation*

$$p_{i,j}(n_1 + n_2) = \sum_{k\in S} p_{i,k}(n_1) p_{k,j}(n_2) \tag{1}$$

holds.

The right hand side of Equation (1) can be considered as a dot product of two vectors, and thus we see that the transition matrix satisfies

$$P^n = \begin{bmatrix} p_{1,1}(n) & \dots & p_{1,m}(n) \\ \vdots & \ddots & \\ p_{m,1}(n) & & p_{m,m}(n) \end{bmatrix}.$$

Therefore, the distribution after n steps is πP^n. Now, to describe the particular models which we focus on, we define the following two terms.

Definition 2.1. A state i is *aperiodic* if $\gcd\{n \geq 1 : p_{i,i}(n) > 0\} = 1$.

Definition 2.2. A chain is *irreducible* if for any two states i, j, there exist $t_1, t_2 > 0$ such that $p_{i,j}(t_1) > 0$ and $p_{j,i}(t_2) > 0$.

It is a well-known result [17] that if a Markov process is irreducible and every state is aperiodic, then there exists a unique distribution λ, known as the equilibrium (or stationary) distribution, such that $\lambda P = \lambda$. The aperiodicity and irreducibility are necessary to avoid in the long run both oscillatory behavior or hidden states, respectively. As such, we assume both always hold. For the remainder of this paper, we focus on Markov State Models, which are models of a system that consider progression in the system to be a Markov process. As such, reference to the *system* actually means the MSM modeling the system.

Now, while each λ is unique given P, a given λ may correspond to multiple P's. To determine the precise transitions of a given system, we consider various *cuts*, which involve inhibiting certain transitions between states and measuring the equilibrium distribution of the altered system. We focus on overdamped systems, so when a transition is cut, say that from state 1 to state 2, then the probability $p_{1,2}(1)$ shifts to the probability of not transitioning. Namely, if the system after the cut has transition probabilities $p'_{i,j}(1)$, then we have $p'_{1,2}(1) = 0$ and $p'_{1,1}(1) = p_{1,1}(1) + p_{1,2}(1)$.

3 Cut Minimization

For efficiency, we seek the minimum number of necessary measurements to uniquely fix a system. Accordingly, we must consider the number of variables and the maximum amount of information contained in each system. We quantify the amount of information per cut by the number of linearly independent equations arising from the equilibrium distribution, or eigenvector with corresponding eigenvalue 1.

Lemma 3.1. *In an n-state system, knowledge of the equilibrium distribution begets at most $n-1$ linearly independent equations, and thus at most $n-1$ new pieces of information.*

Proof. See Appendix A. □

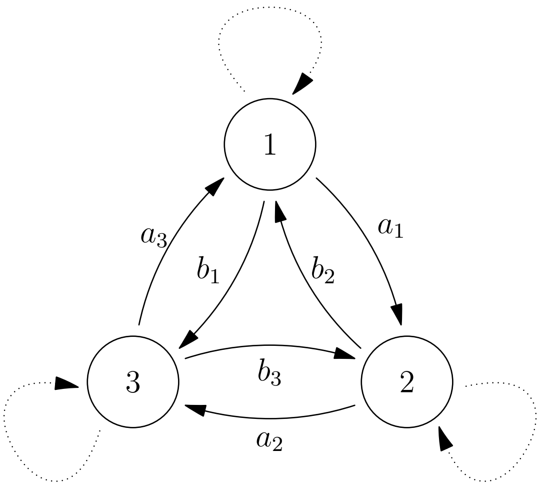


Figure 1: Labeling of 3-state system transition probabilities

We now focus our attention on the three-state case. Simply solving the system of equations which arises from $\lambda P = \lambda$ and referring to Figure 1, we get the unhomogenized equation

$$\lambda = (a_2a_3 + b_2b_3 + b_2a_3, a_1a_3 + b_1b_3 + b_3a_1, a_1a_2 + b_1b_2 + b_1a_2). \tag{2}$$

3.1 Complete Systems

We call a system *complete* if, for all $i \neq j$, we have $p_{i,j}(1) > 0$, so none of the non-diagonal entries in the transition matrix are 0.

Lemma 3.2. *In a three-state system, if blocking off a single transition does not alter the equilibrium distribution, the system is not complete.*

Proof. See Appendix B. □

For the remainder of this subsection, we only consider complete systems, so by Lemma 3.2 equilibrium distributions are changed when deleting an edge. In light of Lemma 3.1, from each measurement we derive at most two equations, and therefore three measurements are necessary to solve for all six probabilities. Now, we consider the coefficient matrix Q for such equations, so the system is $Q \cdot (a_1, a_2, a_3, b_1, b_2, b_3)^T = 0$.

Theorem 3.3 (Minimum Cut Feasibility)**.** *There exist two distinct cuts which, in conjunction with knowledge of the initial equilibrium distribution, are sufficient to uniquely determine a complete, three-state system.*

Proof. We will determine the initial equilibrium distribution and the equilibria defined by deleting two consecutive transitions. Without loss of generality, we cut the transitions from state 1 to state 2 and state 2 to state 3. We use subscripts to denote components of λ, so, for example, Equation (2) gives $\lambda_1 = a_2a_3 + b_2b_3 + b_2a_3$. Now, if P is the transition matrix for the system, from $\lambda(P - I) = 0$, we get

$$(a_2a_3+b_2b_3+b_2a_3, a_1a_3+b_1b_3+b_3a_1, a_1a_2+b_1b_2+b_1a_2) \begin{bmatrix} -a_1 - b_1 & a_1 & b_1 \\ b_2 & -a_2 - b_2 & a_2 \\ a_3 & b_3 & -a_3 - b_3 \end{bmatrix} = 0.$$

Therefore, the first two components of the multiplication give $\lambda_1a_1 + \lambda_1b_1 - \lambda_2b_2 - \lambda_3a_3 = 0$ and $\lambda_1a_1 - \lambda_2a_2 - \lambda_2b_2 + \lambda_3b_3 = 0$. Encoding this in Q, we see the first two rows of Q are

$$Q = \begin{bmatrix} \lambda_1 & 0 & -\lambda_3 & \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & -\lambda_2 & 0 & 0 & -\lambda_2 & \lambda_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Now, we let $\lambda^{(1)}$ be the new equilibrium distribution upon setting $a_1 = 0$, which Equation (2) gives as $(a_2a_3 + b_2b_3 + b_2a_3, b_1b_3, b_1b_2 + b_1a_2)$. If P_1 is the transition matrix for this system,

$$P_1 = \begin{bmatrix} 1-b_1 & 0 & b_1 \\ b_2 & 1-a_2-b_2 & a_2 \\ a_3 & b_3 & 1-a_3-b_3 \end{bmatrix} \implies P_1 - I = \begin{bmatrix} -b_1 & 0 & b_1 \\ b_2 & -a_2-b_2 & a_2 \\ a_3 & b_3 & -a_3-b_3 \end{bmatrix},$$

and then the first two components of $\lambda^{(1)}(P_1 - I) = 0$ are $-\lambda_1^{(1)}b_1 + \lambda_2^{(1)}b_2 + \lambda_3^{(1)}a_3 = 0$ and $\lambda_2^{(1)}a_2 + \lambda_2^{(1)}b_2 - \lambda_3^{(1)}b_3 = 0$. We get analogous equations for $a_2 = 0$, and thus arrive at

$$Q = \begin{bmatrix} \lambda_1 & 0 & -\lambda_3 & \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & -\lambda_2 & 0 & 0 & -\lambda_2 & \lambda_3 \\ 0 & 0 & \lambda_3^{(1)} & -\lambda_1^{(1)} & \lambda_2^{(1)} & 0 \\ 0 & \lambda_2^{(1)} & 0 & 0 & \lambda_2^{(1)} & -\lambda_3^{(1)} \\ \lambda_1^{(2)} & 0 & 0 & 0 & -\lambda_2^{(2)} & \lambda_3^{(2)} \\ 0 & 0 & \lambda_3^{(2)} & -\lambda_1^{(2)} & 0 & \lambda_3^{(2)} \end{bmatrix}. \tag{3}$$

Because we assume the system is complete, we may use the symbolic rank function in MATLAB to find rank$(Q) = 5$, so by the Rank-Nullity Theorem the nullity is 1; namely, there is a unique (up to a single scaling factor, see Section 4) solution to $Q \cdot (a_1, a_2, a_3, b_1, b_2, b_3)^T = 0$, as desired.

□

Corollary 1. *In a complete, three-state system, there exist three measurements which are both necessary and sufficient to uniquely determine the transition matrix and thus dynamics.*

Proof. Necessity comes from Lemma 3.1, and Theorem 3.3 gives sufficiency. □

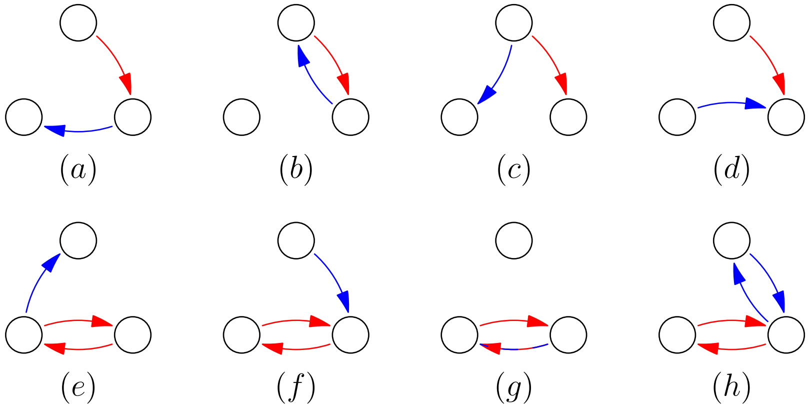


Figure 2: Shown are the eight possible pairs of cuts, with arrows denoting cut transitions. Red denotes transitions blocked in one cut and blue the other. Note that one of the edges in case (g) is cut in both situations and is therefore multichromatic.

Under the physically motivated assumptions that we can only cut a single transition, or possibly a transition and its reverse simultaneously (e.g. the transitions from state 1 to state 2 and state 2 to state 1), there are eight distinct choices of cuts possible in the three-state case, depicted in Figure 2. The proof of Theorem 3.3 considers case (a), and analogously evaluating the coefficient matrices Q for the seven remaining situations, we can similarly utilize MATLAB to determine exactly which of the eight cases uniquely identify the transition probabilities.

Proposition 3.4. *For a complete, three state system, categorizing as in Figure 2, cutting edges as in situations* $(a), (b), (d), (e), (f)$, *and* (g) *all provide sufficient information to determine the system's transition matrix, while cases* (c) *and* (h) *are underdetermined.*

The uniqueness provided by cuts, in particular the consecutive cuts in situation (a), does not appear to be unique to the three-state situation.

Conjecture 3.5. *In a complete, n-state system, n measurements are necessary and sufficient to uniquely determine the transition matrix and thus system dynamics.*

From Lemma 3.1, we see that when there is only one degree of freedom with scaling, there may be up to $n^2 - n - 1$ unknown variables and only $n - 1$ equations arise from each measurement. Consequently, $\left\lceil \frac{n^2-n-1}{n-1} \right\rceil = n$ measurements are necessary, establishing half of the conjecture.

Utilizing the same consecutive cut procedure as in situation (a) of Figure 2, empirical trials in MATLAB provide numerical evidence for the conjecture in four- and five-state systems. If true, this would be a vast improvement in the complete case on any recursive $O(n!)$ algorithm, bringing it to the minimal possible $\Theta(n)$ in the number of cuts.

3.2 Degeneracies

Typically, MSMs will not have direct transitions between every pair of states. This means that, in such degenerate systems, various transition probabilities are 0. Because additional zeros generally decrease matrix rank, or at least have the capacity to invalidate the row operations necessary to perform symbolic Gaussian-elimination, the approaches in dealing with complete systems no longer work. To wrap up the three-state case, we consequently identify four unique degenerate situations in which all three states retain nonzero distributions at equilibrium. Assuming it is known in advance which situation the system is in, we demonstrate what is determinable, labeling cases in the notation of Figure 3.

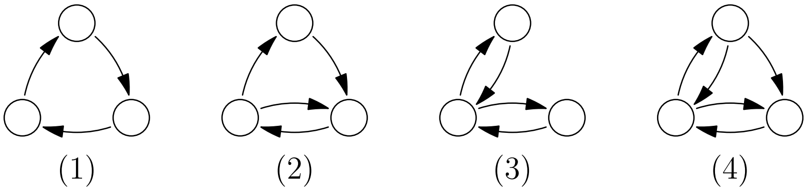


Figure 3: The four, 3-state degenerate cases, with arrows denoting nonzero transitions

Theorem 3.6. *In every degenerate* 3*-state system, less information is necessary to determine the system than in the complete case. In particular, zero cuts suffice in cases* (1) *and* (3)*, while one cut suffices in cases* (2) *and* (4)*.*

Proof. See Appendix C. □

4 Scalability

While previous sections have demonstrated that transition matrices can be mostly recovered given an adequate choice of cuts, there is a underlying hole in the ability to use this information to truly model the dynamics of the systems under consideration. Naturally, system dynamics rely greatly on the time taken for transitions to occur and the system to converge to equilibrium [18–20]. When measurements are only taken once equilibrium has been reached, the rate of this convergence can only be determined relative to other times, not explicitly and independently.

The problem arises from the homogeneity of the equations acquired from the measured equilibria; if the transition probabilities between distinct states are scaled by some constant c, then $Q \cdot (a_1, a_2, a_3, b_1, b_2, b_3)^T = 0 = Q \cdot (ca_1, ca_2, ca_3, cb_1, cb_2, cb_3)^T$, so the system with scaled probabilities has the same equilibrium as the original regardless of which cuts are made. We refer to such degrees of freedom as *unknowable scaling factors*, as they are scaling factors of transition probabilities which do not affect equilibrium and are thus undetectable when we only cut transitions and measure equilibria. While complete systems only have a single scaling factor, some systems have multiple sets of transition probabilities which, when each probability in the set is scaled by the same factor, leave all equilibrium distributions unchanged. These sets function almost independently of each other and therefore act like separate subsystems. The number of such sets, or just the number of such hidden scaling factors, depends on the structure of the system. We analyze different properties of systems that affect the number of unknowable scaling factors so that we may characterize the extent of this scalability problem.

To do so, we utilize the Markov Chain Tree Theorem which establishes and generalizes Equation (2). For a connected, weighted, directed graph $G = (V, E)$, we call a subgraph T a *directed spanning tree* (DST) with root at a vertex v if T is a spanning tree, so it is acyclic and has vertex set V, and there exists a unique directed path from every vertex $u \neq v$ to v. For an example, see the graph in Figure 5 where the grey edges are bidirectional and all edges have unwritten weights. Denote by $\mathcal{T}_v(G)$ the set of DSTs with vertex v as a root. The *weight* $w(T)$ of a directed spanning tree T is the product of the weights of its edges. Now, for an irreducible MSM with transition matrix P, we consider its associated transition graph G_P, the directed graph with P as its adjacency matrix. The relevant result, thought to have been found by Kirchhoff, is that the equilibrium distribution of the system is expressible by considering spanning trees of G_P, and a simple proof is given by Kruckman [21].

Theorem 4.1 (Markov Chain Tree Theorem)**.** *Consider an irreducible, aperiodic n-state MSM with transition matrix P. Label its vertices 1 through n, and let λ represent the equilibrium distribution, with λ_i the ith component. Then,*

$$\lambda_i = \frac{\sum_{T \in \mathcal{T}_i(G)} w(T)}{\sum_{j \in [n]} \sum_{T \in \mathcal{T}_i(G)} w(T)}. \tag{4}$$

Next, we define an *articulation vertex* of a connected graph G as any vertex v such that removing v and its incident edges disconnects G. Furthermore, a *block* of a graph is a maximal connected subgraph which has no articulation vertices (see Figure 4 or the red subgraph in Figure 5). Every graph has a decomposition into such blocks, so denote by $b(G)$ the number of blocks in such a decomposition for a graph G, with component blocks $H_1, H_2, ..., H_{b(G)}$.

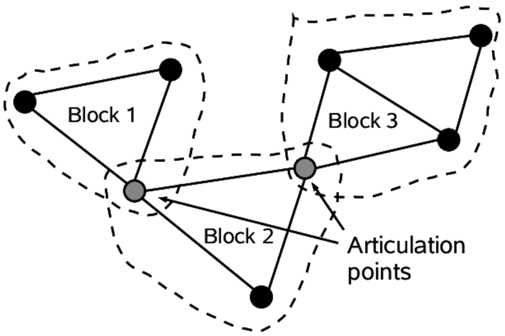


Figure 4: Depiction of a block decomposition with articulation vertices [22]

Lemma 4.2. *In a directed spanning tree $T \in \mathcal{T}_v(G)$, every vertex other than v has out-degree* 1*, while v's out-degree is* 0*.*

Proof. If v had nonzero out-degree, then the edge to its out-neighbor along with its out-neighbor's path to v would form a cycle, contradiction. For a vertex $u \neq v$, the incident edge of its path to v is an outgoing edge of u, so u has out-degree at least 1. If u has two outgoing edges, say to vertices u_1 and u_2, then the paths from u_1, u_2 to v converge either at v or an earlier vertex, forming a cycle when these paths are considered with u, u_1, and u_2. With this contradiction, we have the result. □

It is also well known [23] that every graph has a corresponding block-cutpoint graph, $\mathcal{B}(G)$, a bipartite tree with partition (A, B) where A is the set of articulation vertices, B is the set of blocks condensed into a single vertex, and two vertices a, b are adjacent if the expanded block represented by b contains a.

Lemma 4.3. *For a connected graph G, directed spanning tree $T \in \mathcal{T}_v(G)$ for some vertex v, and block H of graph G, the subgraph $T \cap H$ is a directed spanning tree of H.*

Proof. We split into cases.
Case (1): $v \in H$. If the path of every other vertex $u \in H$ to v remains in H we are done, so assume $\exists u \in H$ such that the path from u to v includes an edge in at least one other block H' sharing an articulation vertex u_1 with H, where u_1 may be u. The path from u must pass through u_1 to go into H', and thus there is some outgoing edge of u_1 in H'. However, consider the path from u_1 to v. If the edge emanating from u_1 is in H then u_1 has an outgoing edge in H and a separate outgoing edge in H', contradicting Lemma 4.2. Hence, the path from u_1 to v begins in H', but it must return to H through some articulation vertex. If this vertex differs from u_1, then we would have a cycle in $\mathcal{B}(G)$, contradicting the fact that the block-cutpoint graph is a tree. Therefore, the path returns to u_1, but then we have found a cycle, contradiction. As such, $T \cap H$ forms a directed spanning tree of H rooted at v.

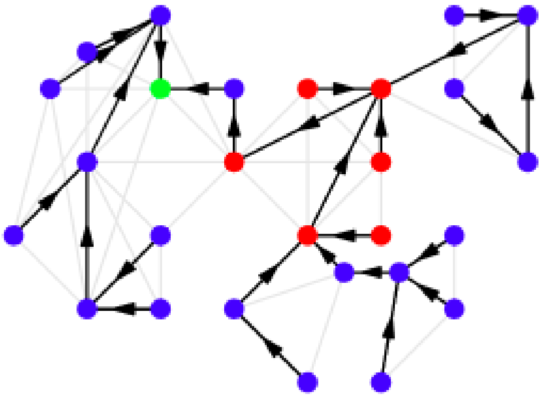


Figure 5: A DST with green vertex as root; the red block's sub-spanning tree is evident

Case (2): $v \notin H$ (see Figure 5). Every path from a vertex u in H to v must pass through the same articulation vertex to leave H and eventually reach a block with v, as if there were two such vertices it would contradict the acyclicity of $\mathcal{B}(G)$. If we call this articulation point u_1, then every path from a u in H to v includes a path from u to u_1. Due to the result from case 1, this path must be completely contained in H, and because this is true for every vertex of H other than u_1, we see that $T \cap H$ is a directed spanning tree of H rooted at u_1. □

Lemma 4.3 indicates that for a spanning tree T of G, you can freely rearrange the edges of $T \cap H$ such that as long as they still form a spanning tree on H, then they, along with the edges of T not in H, still form a spanning tree on G. This freedom to permute the edges of blocks is what creates the scaling factors.

Theorem 4.4. *Consider the transition graph G_P of a given MSM with transition matrix P. Then, only using equilibrium data, the process has at least $b(G_P)$ unknowable scaling factors.*

Proof. Consider a given block H_k; we claim that scaling up the weights of each edge in H_k by some constant c maintains the equilibrium distribution. As the denominator of Equation (4) is a homogenization factor, it is the same for all the λ_i so we may focus solely on the numerator, $\sum_{T \in \mathcal{T}_i(G)} w(T)$. Because G_P has n vertices, each nonzero summand $w(T)$ is the product of $n-1$ distinct edge weights, as either all $n-1$ edges in T have positive weight or one of them is 0 in which case $w(T) = 0$ and is thus irrelevant. Scaling the weight of every edge in H_k by c, every nonzero term in the sum is now multiplied by $c^{T(k)}$, where $T(k)$ represents the number of edges of H_k in the spanning tree T, or equivalently the number of factors of $w(T)$ which are now-scaled weights from edges in H_k.

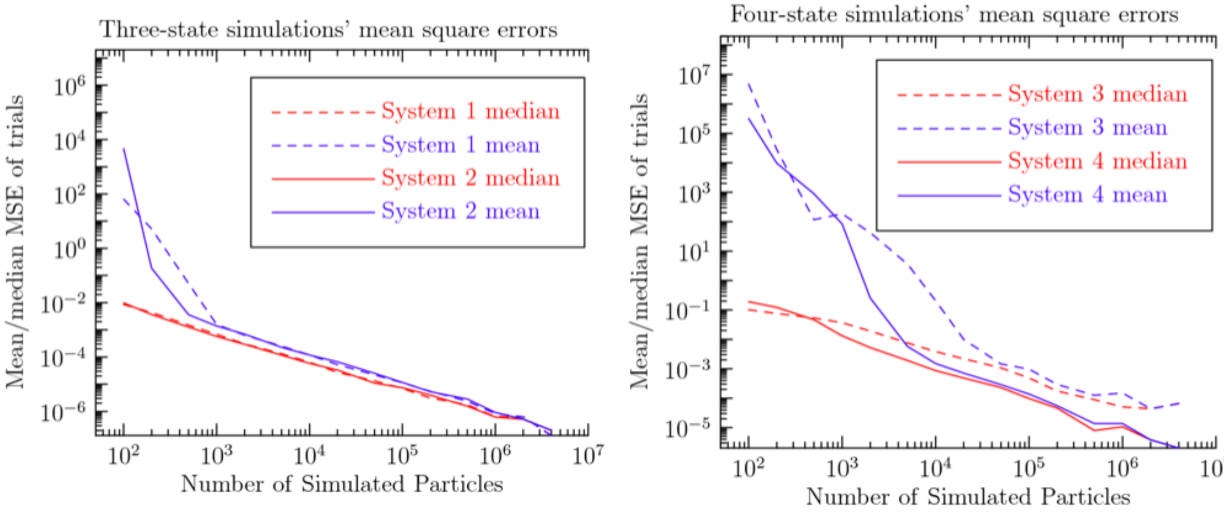
By Lemma 4.3, $T \cap H$ is a directed subtree of H so $T(k)$ is just one less than the number of vertices of H_k. This is independent of T, and thus every nonzero term of the sum over spanning trees of G rooted at i is scaled by $c^{T(k)}$. Therefore, scaling the weights in H_k scales the whole sum by $c^{T(k)}$. Because this is all independent of i, every component of the equilibrium distribution is scaled by this same factor, and therefore after homogenization, the equilibrium distribution remains unchanged. Accordingly, we can independently choose to scale all the edges in any number of the H_i without changing the only measurable component of the system: the equilibrium distribution. Hence, as there are $b(G_P)$ of these H_i, each of which may independently have its edge weights scaled without affecting equilibrium distributions, there are at least $b(G_P)$ unknowable scaling factors. □

5 Verification of Minimum Cut Feasibility

Theoretical understandings of MSMs are important, but while our Minimum Cut Feasibility theorem (3.3) proves that the initial distribution and two distinct cuts can in theory determine a system's transition probabilities, there is no confirmation of a practical approach fit for actual modeling. The finiteness of systems and subsequent noise give reason to question the effectiveness of purely theoretical results. Accordingly, we set out to demonstrate that the sufficiency in our Minimum Cut Feasibility theorem (3.3) translates to a reasonably successful algorithm by running simulations in MATLAB.

In each simulation of an N particle or cell system, we arbitrarily pick transition probabilities to describe a system and determine the equilibrium distributions for that system and the system with certain transitions cut. We acquire experimental equilibria by partitioning the interval $(0, 1)$ into subintervals with lengths proportional to the components of the actual equilibria, then randomly sampling N numbers from $(0, 1)$ and assigning each to the corresponding state. We use the experimental distributions to determine the coefficient matrix Q. However, finding the experimental transition probabilities requires the nullspace of Q, yet Q is almost always full rank when using the random error-ridden experimental data. As such, we instead use low-rank approximation and find the nullspace of Q^*, the truncated singular value decomposition of Q as described by the Eckart-Young-Mirsky Theorem [24], which guarantees Q^* as the matrix of desired rank with closest distance to Q under the Frobenius norm.

To compare the experimental probabilities from the nullspace of Q^* with the actual values, we homogenize so the sum of all probabilities is 1 and then evaluate the mean squared error (MSE) across the $n^2 - n$ probabilities in the n-state case. The number of trials run for each number of particles N is $\min\left(10000, \frac{4000000}{N}\right)$. We conduct simulations four separate times with different, quasirandom equilibrium distributions, twice with three-states and twice with four-states. The equilibrium data corresponds to the full system's equilibrium in addition to consecutive cuts; we cut the transitions from state 1 to state 2 and state 2 to state 3, as well as state 3 to state 4 in the four-state trials. The data is shown in Figure 6, where both the mean MSE and median MSE over all trials for a given system and N are graphed. Graph (a) corresponds to the two 3-state systems, and graph (b) to the two 4-state systems. The simulation results demonstrate relatively small errors, orders of magnitude smaller than the probabilities in the systems. The linearity in the log-log graph is apparent, although there is noise for the smallest and largest N especially in (b). The larger number of fluctuations for small N is attributable to very high means due to outlier situations in which there may have been negative values calculated from Q^*s nullspace. These induce a sum of the components



(a)

(b)

Figure 6: Simulation data for three- and four-state systems

which is close to zero, which upon homogenization greatly scales the data, causing high errors. Since the median is affected little by such outliers, it better reflects the true error in these simulations. Additionally, the noise for the largest N comes from running at most 4 trials due to computational inefficiency and runtime constraints. Finally, the proximities of the median graphs' slopes to -1 suggest standard deviations on the order of $\frac{1}{\sqrt{N}}$ upon transforming back from the doubly logarithmic scale.

6 Conclusion

In this paper, we focus on the problem of determining MSM dynamics when only equilibrium data is available. This limited availability of data is a very real hurdle in situations where the most feasible measurements are disruptive. We consider the effect of cutting, or inhibiting, different transitions on equilibrium distributions, using information from the changes to uniquely characterize the system. We first show that in complete, three-state, overdamped systems, two cuts are both necessary and sufficient to reconstruct the transition matrix. Numerical evidence suggests the minimum necessary number being sufficient generalizes to any number of states. We also determine the number of blocks in the transition graph as a lower bound on the number of unknowable scaling factors of the system, or degrees of freedom to scale transition probabilities without changing any equilibrium distributions. Finally, to test applicability, we simulate complete three- and four-state systems and demonstrate that the sufficient algorithm we present is practical in that it determines transitions probabilities from experimental data with an error that decreases linearly in the number of particles considered. In future research, we intend to refine the simulated approximation by considering different rank-reduction techniques and also prove our conjectured generalization of minimum cut sufficiency.

7 Acknowledgments

I would like to thank my mentor Dominic Skinner of the MIT Math Department for proposing the subject of this paper and guiding me along with advice and ideas throughout the research process. Thanks also to Dr. Tanya Khovanova of the MIT Math Department for providing advice and coordinating this research. Thanks to Dr. John Rickert of the Rose-Hulman Institute of Technology for providing excellent feedback on this paper and counsel regarding the research process. Thank you to Dr. Linda Westrick of Penn State University for her crucial comments and suggestions about this paper. Thank you to Prof. David Jerison and Prof. Ankur Moitra of the MIT Math Department for organizing this research. Thank you to Sean Elliott, AnaMaria Perez, Assylbek Olzhabayev, Rupert Li and Jason Liu for proofreading and providing paper advice, as well as to Shashvat Shrivastava for formatting and Latex help. Finally, thank you to the MIT Math Department, RSI, and CEE for giving me this opportunity, as well as the Department of Defense, the Alan Kane Charitable Trust, and the Kenneth and Myra Monfort Charitable Foundation for funding this research.

References

[1] J. Pitman and M. Yor. A guide to Brownian motion and related stochastic processes. 02 2018.

[2] N. Gō. Protein folding as a stochastic process. *Journal of Statistical Physics*, 30(2):413–423, Feb 1983.

[3] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 10 2008.

[4] T. Lipniacki, P. Paszek, A. Marciniak-Czochra, A. Brasier, and M. Kimmel. Transcriptional stochasticity in gene expression. *Journal of Theoretical Biology*, 238(2):348–367, 01 2006.

[5] P. Pearce, F. G. Woodhouse, A. Forrow, A. Kelly, H. Kusumaatmaja, and J. Dunkel. Learning dynamical information from static protein and sequencing data. *bioRxiv*, 2019.

[6] B. Husic and V. Pande. Markov State Models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 02 2018.

[7] D. Shukla, C. X. Hernández, J. K. Weber, and V. S. Pande. Markov State Models provide insights into dynamic modulation of protein function. *Accounts of Chemical Research*, 48(2):414–422, 2015. PMID: 25625937.

[8] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *Journal of Chemical Physics*, 134(17):174105, 05 2011.

[9] N. Deng, W. Dai, and R. Levy. How kinetics within the unfolded state affects protein folding: an analysis based on Markov State Models and an ultra-long md trajectory. *Journal of Physical Chemistry B*, 117(42):12787–99, 10 2013.

[10] F. Noé and S. Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Structural Biology*, 18(2):154–162, 04 2008.

[11] J. Chodera and F. Noé. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Structural Biology*, 25:135–144, 04 2014.

[12] B. K. Chu, M. J. Tse, R. R. Sato, and E. L. Read. Markov State Models of gene regulatory networks. *BMC Systems Biology*, 11(14), 02 2017.

[13] S. Kim, E. R Dougherty, Y. Chen, M. Bittner, and E. Suh. Can Markov Chain Models mimic biological regulation? *Journal of Biological Systems*, 10, 03 2003.

[14] A. Rosenberg, C. M. Roco, R. A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. Gray, D. J. Peeler, S. Mukherjee, W. Chen, S. Pun, D. Sellers, B. Tasic, and G. Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360:eaam8999, 03 2018.

[15] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[16] B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé. Estimation and uncertainty of reversible Markov models. *Journal of Chemical Physics*, 143:174101, 11 2015.

[17] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, 2005.

[18] P. Clote. Protein folding, the Levinthal paradox and rapidly mixing Markov chains. volume 1644, pages 701–702, 01 1999.

[19] E. Suárez, J. L. Adelman, and D. M. Zuckerman. Accurate estimation of protein folding and unfolding times: Beyond Markov State Models. *Journal of Chemical Theory and Computation*, 12(8):3473–3481, 2016. PMID: 27340835.

[20] P. Robert. Mathematical Models of Gene Expression. *arXiv e-prints*, page arXiv:1905.02578, May 2019.

[21] A. Kruckman, A. Greenwald, and J. R. Wicks. An elementary proof of the Markov chain tree theorem. 2010.

[22] S. Jansen, S. Tate, D. Tsagkarogiannis, and D. Ueltschi. Multispecies virial expansions. *Communications in Mathematical Physics*, 330, 04 2013.

[23] F. Harary and G. Prins. The block-cutpoint-tree of a graph. *Publicationes Mathematicae Debrecen*, 13:103–107, 1966.

[24] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sep 1936.

A Proof of Lemma 3.1

Lemma. *In an n-state system, knowledge of the equilibrium distribution begets at most $n-1$ linearly independent equations, and thus at most $n-1$ new pieces of information.*

Proof. Firstly, as the equilibrium distribution λ is merely a left eigenvector with corresponding eigenvalue 1, the only condition on λ is that $\lambda(P-I)=0$, and because the left side is of dimension n, there are at most n equations. However, the final equation, or the last component of $\lambda(P-I)$, is merely the sum of the first $n-1$. Indeed, consider the ith component of each such equation. Denote $\lambda=(\lambda_1,\lambda_2,...,\lambda_n)$, so the ith component of the jth equation is $\lambda_i p_{i,j}(1)$, for $i \neq j$. When $i=j$, because P is stochastic, the (i,i)th element of P is $1-\sum_{k=1,k\neq i}^{n} p_{i,k}(1)$ so the (i,i)th element of $P-I$ is $-\sum_{k=1,k\neq i}^{n} p_{i,k}(1)$, and the ith component of the ith equation is $\lambda_i\left(-\sum_{k=1,k\neq i}^{n} p_{i,k}(1)\right)$. Summing these ith components over j, we get

$$\begin{aligned}\sum_{j=1}^{n}\lambda_i p_{i,j}(1) &= \lambda_i\sum_{j=1}^{n} p_{i,j}(1)\\ &= \lambda_i\left(\sum_{j=1,j\neq i}^{n} p_{i,j}(1)+(p_{i,i}(1)-1)\right).\end{aligned}$$

Finally, the last expression is $\lambda_i\left(\sum_{j=1,j\neq i}^{n} p_{i,j}(1)+\left(-\sum_{k=1,k\neq i}^{n} p_{i,k}(1)\right)\right)=0$. Therefore, summing all n linear equations gives the tautology $0=0$, and therefore the last equation is the sum of the negations of the first $n-1$ equations. This means it can be expressed as a linear combination of them and is not linearly independent, implying there are at most $n-1$ linearly independent equations. □

B Proof of Lemma 3.2

Lemma. *In a three-state system, if blocking off a single transition does not alter the equilibrium distribution, the system is not complete.*

Proof. Assume, for the sake of contradiction, that there exists some complete system with transition matrix P, and some transition probability $p_{i,j}(1) > 0$ which, when set equal to 0, leaves λ unchanged. Without loss of generality, suppose we change $p_{1,2}(1)$, so a_1, to 0. Setting $a_1 = 0$ in Equation (2), we get $\lambda = (a_2a_3 + b_2b_3 + b_2a_3, b_1b_3, b_1b_2 + b_1a_2)$. Now, because λ_1 remains unchanged, and the first components are identically equal, we must have the second components are equal. Therefore, $b_1b_3 = a_1a_3 + b_1b_3 + b_3a_1$, which equates to $a_1(a_3 + b_3) = 0$. This is not possible because the system is complete, so $a_1, a_3, b_3 > 0$. Thus, we reach a contradiction, so the system is not complete. □

C Proof of Theorem 3.6

Theorem. *In every degenerate* 3*-state system, less information is necessary to determine the system than in the complete case. In particular, looking at Figure 7 we find zero cuts suffice in cases* (1) *and* (3)*, while one cut suffices in cases* (2) *and* (4)*.*

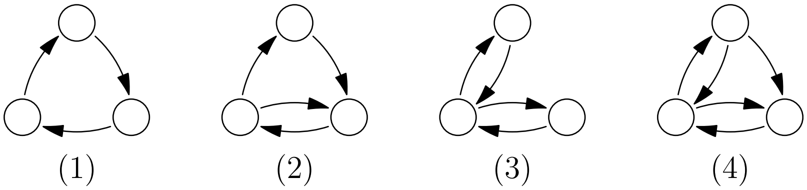


Figure 7: The four, 3-state degenerate cases, with arrows denoting nonzero transitions

Proof. We label cases in the notation of Figure 7 and referring to the same orientation and transitions as in Figure 1.

Case (1): Substituting $b_1 = b_2 = b_3 = 0$ into Equation (2), we find $\lambda = (a_2a_3, a_1a_3, a_1a_2)$, which by scaling is equivalent to $\left(\frac{1}{a_1}, \frac{1}{a_2}, \frac{1}{a_3}\right)$. As such, reciprocating the measured equilibrium is sufficient, so no cuts are necessary, only knowledge of the initial equilibrium distribution.

Case (2): From Equation (2), substituting $b_1 = b_2 = 0$ we get $\lambda = (a_2a_3, a_1a_3 + b_3a_1, a_1a_2)$, and thus we know $\frac{a_1}{a_3}$ and $\frac{a_3+b_3}{a_2}$. While this is insufficient, cutting b_3 gives case (1), from which we determine a_1, a_2, a_3 up to scaling. Therefore, since we know $\frac{a_3+b_3}{a_2}$, we can determine b_3, so the initial system and a single cut suffice.

Case (3): Setting $a_1 = b_2 = 0$ into Equation (2) gives $\lambda = (a_2a_3, b_1b_3, b_1a_2)$. Thus, we can determine $\frac{a_2}{b_3}$ and $\frac{b_1}{a_3}$. Checking all other cuts, we can't gain any more information, a fact which will later follow from Theorem 4.4. Therefore, all information that can be determined by cuts is recoverable from the initial equilibrium.

Case (4): Setting $b_2 = 0$ into Equation (2), we get $\lambda = (a_2a_3, a_1a_3 + b_1b_3 + b_3a_1, a_1a_2 + b_1a_2)$. Now, cutting the top right, from case (3) we know $c_1 := \frac{a_3}{b_1}$ and $c_2 := \frac{a_2}{b_3}$, so the known initial equilibrium becomes $(a_2a_3, a_1a_3 + c_1c_2a_2a_3 + c_2a_2a_1, a_1a_2 + c_1a_3a_2)$. The ratio of the third to the first component gives us $\frac{a_1}{a_3}$, and the ratio of the second to the first component gives $\frac{a_1}{a_2}$, which completes the system. Thus, the initial system and a single cut suffice. □