53rd IEEE Conference on Decision and Control
December 15-17, 2014. Los Angeles, California, USA

Aggregation of Markov Chains: an Analysis of Deterministic Annealing Based Methods

Yunwen Xu, Carolyn L. Beck, Srinivasa M. Salapaka

***Abstract*— We develop a method for aggregating large Markov chains into smaller *representative* Markov chains, where Markov chains are viewed as weighted directed graphs, and *similar* nodes (and edges) are aggregated using a *deterministic annealing* approach. The notions of *representativeness* of the aggregated graphs and *similarity* between nodes in graphs are based on a newly proposed metric that quantifies connectivity in the underlying graph. Namely, we develop notions of distance between subchains in Markov chains, and provide easily verifiable conditions that determine if a given Markov chain is nearly decomposable, that is, conditions for which the deterministic annealing approach can be used to identify subchains with high probability. We show that the aggregated Markov chain preserves certain *dynamics* of the original chain. In particular we provide explicit bounds on the ℓ_1 norm of the error between the aggregated stationary distribution of the original Markov chain and the stationary distribution of the aggregated Markov chain, which extends on longstanding foundational results (Simon and Ando, 1961).**

I. Introduction

Data driven and empirical methods used to study complex systems result in large and unwieldy models, for example in applications ranging from network analysis [1], to economics [2], [3], to neuroscience [4], [5]. In many cases the models comprise large graphs and networks, such as connection structures in social networks [6], metabolic networks [7], and in brain activity maps represented by Markov chains [8]. For tractable analysis and design, succinct models are necessary. Further, for large graph models, the identification of underlying coarse connectivity structure is frequently of primary interest. In this context, clustering and aggregation methods play an important role in terms of both *tractability*, for example in aggregation of large Markov chain models, and *identification* of underlying network structures.

Recently, clustering based algorithms have been proposed for the purpose of determining reduced dimension graph-based models [9]. As such, an optimal aggregation of nodes in the graph is sought, where the optimality is evaluated based on a distance measure quantifying similarity in connectivity. These algorithms, which have their foundation in the deterministic annealing method proposed by Rose for the vector quantization problem [10], are directly applicable to the Markov chain reduction, or aggregation problem. In this paper, we present analytical results of graph aggregation methods specific to Markov chains; these directly generalize to weighted directed graphs.

Aggregation of Markov chains, or more generally, stochastic systems, has been studied for over 50 years. One of the more common methods for reducing the size of Markov chains is to aggregate the states into superstates based on the concepts of *completely decomposable* and *nearly completely decomposable* subsystems, as originally introduced for stochastic systems in [2] and [11]. In short, a Markov chain or stochastic system is said to be (i) completely decomposable if its associated transition matrix, $\mathbf{P}$, can be permuted to a strictly block diagonal form; and (ii) nearly completely decomposable (NCD) if $\mathbf{P}$ can be partitioned (possibly following permutations) such that transitions occur much more frequently between states within a partition, than across partitions. These partitions are used to form the aforementioned superstates, and can be viewed as comprising *subchains* within the larger Markov chain. In this framework, aggregated states represented by superstates have dynamics which can be viewed as evolving along a similar short-run time scale. Alternatively, interactions between the superstates can be viewed as evolving along a long-run time scale. Systems with the NCD structure occur commonly in a variety of areas, including power systems [12], and economics [2] as a couple examples.

Following the work of Simon and Ando, Courtois [2], [3] showed that the accuracy of the aggregation based approximations proposed in [2] could be analyzed as a function of the *maximum degree of coupling* between the superstates. In [12], a singular perturbation analysis perspective on the aggregation process was provided; in [13] a simulation-based aggregation approach was presented.

In this paper, we analyze aggregation algorithms previously proposed in [9], [14], [15] for reducing the size of directed weighted graphs and Markov chains. These algorithms are applied using an original formulation for a dissimilarity or distance function that explicitly measures how close two graphs or chains are, as determined by the connectivity between nodes or states. The main results of this paper are:

1. Quantification of the effectiveness of deterministic annealing (DA) based methods for *identifying* NCD partitions in Markov chains, based directly on the maximum degree of coupling. We show that our DA-based algorithms correctly identify subchains when the distance between any two states within the same subchain compared to the distance between any two states from different subchains satisfies specific bounds given in terms of this coupling parameter.
2. Statistical quantification of misclassification errors resulting from the DA-based algorithms. We show that the probability of misclassification in our methods can be given explicitly in terms of a Chi-squared distribution.
3. Demonstration of preservation of dynamic properties in

This work was partially supported by the NSF grant CMMI - 1100257.

Yunwen Xu, S. Salapaka and C. L. Beck are from University of Illinois at Urbana-Champaign, Urbana, Illinois. Email: [xu27, salapaka, beck3]@illinois.edu

aggregated chains. We show that even though the aggregation of Markov chains is obtained from a *data-centric* point of view without explicit regard to dynamic behavior, the aggregated chain preserves the stationary distribution. In particular, we show the ℓ_1 norm of the error between the aggregated stationary distribution of a nearly decomposable Markov chain, without loss of generality in the form $\mathbf{P}^*+\epsilon\mathbf{C}$ with $\mathbf{P}^*$ being completely decomposable, and the stationary distribution of the aggregated Markov chain is $O(\epsilon)$. We require no a priori knowledge of $\mathbf{P}^*$ and thus extend the seminal results in [2].

II. Notation

For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote its i^{th} row by $\mathbf{a}(i) \in \mathbb{R}^n$. The vector $\mathbf{e}_i \in \mathbb{R}^n$ represents the i^{th} column of the identity matrix $\mathbf{I}_n \in \mathbb{R}^{n\times n}$; $\mathbf{1}_n = [1,\dots,1]^T$ in $\mathbb{R}^n$. A *partition function* $\phi : \mathcal{N} := \{1,2,\dots,n\} \to \mathcal{M} := \{1,\dots,m\}$ is an *onto* map such that $\phi^{-1}(\mathcal{M})$ is a partition of $\mathcal{N}$. We associate an *aggregation matrix* $\Phi \in \{0,1\}^{n\times m}$, whose ij^{th} entry is $\Phi_{ij} = [\mathbf{\Phi}]_{ij} = 1$ when $\phi(i) = j$ and is otherwise equal to 0. We note that $\mathbf{\Phi}\mathbf{1}_m = \mathbf{1}_n$, and the k^{th} column of $\mathbf{\Phi}$ equals $\sum_{i\in\phi^{-1}(k)} \mathbf{e}_i$.

Markov chains: We adopt standard notation from the literature and specifically from [2], [3], [16] for *completely decomposable (CD)* and *nearly completely decomposable (NCD) Markov chains*. A CD Markov chain can be aggregated to several non-communicating subchains, that is, the transition matrix $\mathbf{P}^*$ of a CD Markov chain is a block-diagonal stochastic matrix $\mathbf{P}^* = diag(\mathbf{P}_1^*,\dots,\mathbf{P}_I^*,\dots,\mathbf{P}_N^*)$ where $\mathbf{P}_I^* \in \mathbb{R}^{n_I\times n_I}$ is a stochastic matrix for all $1 \le I \le N$. The transition matrix $\mathbf{P}$ of an NCD Markov chain may be assumed to be in the form (modulo permutations)

$$\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{n\times n}$ is a bounded matrix (in $\|\cdot\|_1$) and $\epsilon > 0$ is a positive real number. The block structure of $\mathbf{P}^*$ provides a natural partition on the state space $\mathcal{N}$, where the states associated with the K^{th} diagonal block are in the K^{th} subchain, and are indexed as

$$\phi^{-1}(K) = \{\sum_{I=1}^{K-1} n_I + 1,\dots,\sum_{I=1}^{K} n_I\}, \quad 1 \le K \le N.$$

Thus $n = \sum_{I=1}^{N} n_I$. A completely decomposable Markov chain can be aggregated accordingly, where $\mathbf{P}_I^*$ is the transition matrix for the I^{th} subchain, specifically, the block structure of $\mathbf{P}^*$ determines the block structures of both $\mathbf{P}$ and $\mathbf{C}$. We denote the submatrix of $\mathbf{P}$ consisting of the rows in $\phi^{-1}(I)$ and columns in $\phi^{-1}(J)$ by $\mathbf{P}_{IJ}$, and the ab^{th} entry of $\mathbf{P}_{IJ}$ by $P_{a_I b_J}$. To ensure the stochasticity of the resulting $\mathbf{P}$ matrix, entries of $\mathbf{C}$ need to satisfy the following constraints (similar to those defined in [3])

$$\begin{cases} \sum_{k=1}^{n_I} C_{i_I k_I} = -\sum_{J\neq I}\sum_{j=1}^{n_J} C_{i_I j_J}, \\ C_{i_I j_J} \ge 0, \quad 1 \le i \le n_I, \forall j, 1 \le I, J \le N, I \neq J. \end{cases} \tag{2}$$

From these constraints, we see that the diagonal blocks of $\mathbf{C}$ must contain negative entries; thus ϵ has an upper bound under which non-negativity of $\mathbf{P}$ is guaranteed. We also assume each stochastic submatrix $\mathbf{P}_I^*$ corresponds to an *aperiodic* and *irreducible* Markov chain, and so has a *unique* stationary distribution $\boldsymbol{\pi}^T(\mathbf{P}_\mathbf{I}^*) \in \mathbb{R}_+^{n_I}$. Let the stationary distribution of the original Markov chain be $\boldsymbol{\pi}^T(\mathbf{P}) \in \mathbb{R}^n$.

III. State Aggregation via Graph Clustering

A. Optimization formulation

In prior work, [9], [14], [15], we've studied the problem of graph clustering via aggregating nodes with similar edge connections. More specifically, for a general *weighted directed graph* we construct a reduced-order *representative graph*, such that a well-motivated *dissimilarity measure* between the two graphs is minimized. In this paper, we focus on analysis of the dynamic properties that are preserved by the reduced Markov chain when the aggregation algorithm is applied to the original Markov chain, and discuss additional meaningful properties of the reduced chain.

For a given a Markov chain $\mathcal{X}$ with n states whose transition probability matrix is $\mathbf{P}$ we would like to find a low-order Markov chain $\mathcal{Y}$ with m states and transition matrix $\mathbf{Q}$ such that the *dissimilarity* between $\mathcal{X}$ and $\mathcal{Y}$ is minimized; this *dissimilarity* is given by [9], [14], [15]

$$\nu(\mathbf{P},\mathbf{Q}) \stackrel{\triangle}{=} \min_{\mathbf{\Phi},\mathbf{Z}|\mathbf{Q}=\mathbf{Z}\Phi} \sum_{i=1}^{n} \mu_i\, d(\mathbf{p}(i), \mathbf{z}(\phi(i))), \tag{3}$$

where $\{\mu_i\}$ are node weights, ϕ is the partition function corresponding to $\mathbf{\Phi}$, and $d(\cdot,\cdot)$ is a vector distance measure. Typically, $\boldsymbol{\mu}$ is chosen based on prior knowledge about the states. Thus the dissimilarity is viewed as the minimum weighted average distance that can be achieved between the rows $\mathbf{p}(i)$ and rows $\mathbf{z}(\phi(i))$ under the constraint $\mathbf{Z}\mathbf{\Phi} = \mathbf{Q}$, $\mathbf{Z} \in \mathbb{R}_+^{m\times n}$. Recall the row vectors $\mathbf{p}(i)$ and $\mathbf{z}(\phi(i))$ define the outgoing transitions for the i^{th} state of $\mathcal{X}$ and the corresponding state of $\mathcal{Y}$, specified by state partition $\mathbf{\Phi}$. The problem of determining an optimal reduced-order Markov chain $\mathcal{Y}$ is converted to that of solving for a set $\{\mathbf{Q}^\star, \mathbf{\Phi}^\star, \mathbf{Z}^\star\}$ that achieves the minimum value $\nu(\mathbf{P},\mathbf{Q}^\star)$. This problem requires the solution of a combinatorial optimization problem with decision variables being both discrete and continuous valued ($\mathbf{\Phi}$ and $\mathbf{Z}$), which is known to be NP-hard [17]. In [9], [15] we provide a two-step procedure that efficiently solves a relaxed version of this problem, leading to a meaningful approximation.

B. A soft aggregation algorithm

We relax the mixed-integer constraint $\mathbf{\Phi} \in \{0,1\}^{m\times n}$ to a continuous weighting matrix constraint $\tilde{\mathbf{\Phi}} \in [0,1]^{m\times n}$, giving the following *soft clustering problem:*

$$\arg\min_{\mathbf{Q}} \tilde{\nu}(\mathbf{P},\mathbf{Q}) = \arg\min_{\tilde{\mathbf{\Phi}},\mathbf{Z}} \tilde{\rho}_{\tilde{\mathbf{\Phi}},\mathbf{Z}}(\mathbf{P},\mathbf{Q}) \tag{4}$$

$$\text{s.t.} \quad \begin{cases} \tilde{\rho}_{\tilde{\mathbf{\Phi}},\mathbf{Z}}(\mathbf{P},\mathbf{Q}) = trace(\mathbf{\Lambda}\mathbf{D}\tilde{\mathbf{\Phi}}^T) \\ \mathbf{\Lambda} = \text{diag}\,(\boldsymbol{\mu}) \\ D_{ij} = [\mathbf{D}]_{ij} = d(\mathbf{p}(i), \mathbf{Z}(j)) \\ \tilde{\mathbf{\Phi}} \in [0,1]^{n\times m}, \tilde{\mathbf{\Phi}}\mathbf{1}_n = \mathbf{1}_n \\ \mathbf{Q} = \mathbf{Z}\tilde{\mathbf{\Phi}}, \mathbf{Z}\mathbf{1}_n = \mathbf{1}_m, \mathbf{Z} \ge 0. \end{cases}$$

In this formulation, the *soft* aggregation matrix $\tilde{\mathbf{\Phi}}$ defines a non-unique association between a state in $\mathcal{N}$ and a superstate in $\mathcal{M}$. More specifically, the ij^{th} entry of $\tilde{\mathbf{\Phi}}$, $\tilde{\Phi}_{ij}$ represents the association level between the i^{th} state of $\mathcal{N}$ and the j^{th} state of $\mathcal{M}$. The resulting *soft* dissimilarity functions are given by $\tilde{\nu}(\mathbf{P},\mathbf{Q})$ in (4).

The effective "randomness" of these associations $\{\tilde{\Phi}_{ij}\}$ can be quantified by a Shannon entropy term, e.g.,

$H(\mathbf{Q}|\mathbf{P}) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{\Phi}_{ij}\log\tilde{\Phi}_{ij}$. We adapt the DA algorithm to solve a series of entropy-constrained minimization problems (4) with additional constraints $H(\mathbf{Q}|\mathbf{P}) = H_k$ with iteration index k. These constraints enter the problem as a Lagrangian term, $-\beta_k(H(\mathbf{Q}|\mathbf{P}) - H_k)$, with β_k being the Lagrangian multiplier. A series of decreasing values of $H_k(\in [0, \log m])$ may be used, resulting in an *annealing* process, during which the soft aggregation matrix $\tilde{\mathbf{\Phi}}$ approaches a binary-valued *hard* aggregation matrix. In implementation, this is realized more simply by using an increasing (geometric) sequence for β_k. See [9] for more details.

IV. Performance Analysis

A. Subchain identification in Markov chains

Extensive simulations demonstrate that our DA-based algorithm easily identifies subchains in CD as well as NCD Markov chains with transition matrices of the form $\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}$. Here we derive *analytical* conditions on the perturbation $\epsilon\mathbf{C}$ that guarantee separability of subchains, supporting our simulation results.

Assumption *1:* All entries of $\mathbf{C}$ in the *diagonal* blocks are non-positive, i.e., $C_{i_I k_I} \leq 0, \forall i, k, I$.

Definition 1 (Maximum degree of coupling [3]): The maximum degree of coupling between any two partitions is given by

$$\delta \triangleq \max_I \left(\sum_{J\neq I}\sum_{j=1}^{n_J} C_{i_I j_J} \right). \tag{5}$$

Definition 2 (Index of unbalance): For an NCD Markov chain in the form of (1) with completely decomposable component $\mathbf{P}^* = \mathrm{diag}\{\mathbf{P}_I^*\}$, the index of unbalance is

$$\eta \triangleq \max_I \max_{i,j} \|\mathbf{p}^*(i_I) - \mathbf{p}^*(j_I)\|_1. \tag{6}$$

Definition 3 (Separability between subchains): For an NCD Markov chain with states $\mathcal{N}$ and transition matrix $\mathbf{P}$ in the form of (1) with completely decomposable component $\mathbf{P}^* = \mathrm{diag}\{\mathbf{P}_I^*\}$, let $\mathcal{X}_I$ be the states associated with the I^{th} diagonal block of $\mathbf{P}^*$. We say the I^{th} subchain and the J^{th} subchain are *separable* if there exists a separating hyperplane such that the outgoing vectors from every state in $\mathcal{N}_I$ and the outgoing vectors from every state in $\mathcal{N}_J$ do not cross the hyperplane.

Theorem *1:* Suppose the NCD Markov chain $\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}$ has a (CD) nominal transition matrix $\mathbf{P}^*$, and perturbation matrix $\mathbf{C}$ satisfying *Assumption 1*. Further assume the *index of unbalance* for $\mathbf{P}$ for each subchain $\mathbf{P}_I^*$ of $\mathbf{P}^*$ is upper bounded by η. Then, if the perturbation factor satisfies $\epsilon < \frac{1-\eta}{6\delta}$, separability between any two subchains is guaranteed.

Proof: First consider pairwise distances of outgoing vectors from the same subchain. For any two states i_I, j_I associated with the I^{th} subchain, the l_1 distance between their outgoing vectors has upper bound

$$\begin{aligned} d_{in} &\leq \|\mathbf{p}^*(i_I) - \mathbf{p}^*(j_I)\|_1 + \epsilon\|\mathbf{c}(i_I) - \mathbf{c}(j_I)\|_1 \\ &\leq \eta + \sum_{J=1}^{N}\sum_{k=1}^{n_J}\left[|C_{i_I k_J}| + |C_{j_I k_J}|\right] \;\leq\; \eta + 4\epsilon\delta. \end{aligned} \tag{7}$$

Since by assumption all entries in the diagonal blocks of perturbation matrix $\mathbf{C}$ are non-positive, we have $\sum_{k=1}^{n_I}|C_{i_I k_I}| = \sum_{J\neq I}\sum_{j=1}^{n_J} C_{i_I j_J} \leq \delta$. Similarly, for two states i_I and j_J from different subchains, we have

$$\begin{aligned} \|\mathbf{p}(i_I) - \mathbf{p}(j_J)\|_1 = &\underbrace{\sum_{k=1}^{n_I}|P^*_{i_I k_I} + \epsilon(C_{i_I k_I} - C_{j_J k_I})|}_{(a)} \\ &+ \underbrace{\sum_{k=1}^{n_J}|P^*_{j_J k_J} + \epsilon(C_{j_J k_J} - C_{i_I k_J})|}_{(b)} + \underbrace{\epsilon\sum_{N\neq J,I}\sum_{k=1}^{n_N}|C_{i_I k_N} - C_{j_J k_N}|}_{(c)}. \end{aligned} \tag{8}$$

The term $(a) \geq 1 - 2\delta\epsilon$, since $(a) \geq$

$$\sum_{k=1}^{n_I}\left(|P^*_{i_I k_I}| - \epsilon|C_{i_I k_I} - C_{j_J k_I}|\right) = 1 - \epsilon\sum_{k=1}^{n_I}\left(|C_{i_I k_I}| + |C_{j_J k_I}|\right).$$

Similarly, term $(b) \geq 1 - 2\delta\epsilon$, and term $(c) \geq 0$. Therefore, we conclude the following upper and lower bounds for within, and between partition distances,

$$\begin{aligned} d_{in} &\triangleq \max_{1\leq I\leq N}\max_{1\leq i,j\leq n_I}\|\mathbf{p}(i_I) - \mathbf{p}(j_I)\|_1 \leq \eta + 4\delta\epsilon, \text{ and} \\ d_{out} &\triangleq \min_{1\leq I,J\leq N}\min_{1\leq i\leq n_I, 1\leq j\leq n_J}\|\mathbf{p}(i_I) - \mathbf{p}(j_J)\|_1 \geq 2 - 4\delta\epsilon. \end{aligned}$$

The distance between any outgoing vector of subchain $\mathcal{X}_I$ and any outgoing vector of subchain $\mathcal{X}_J$ is at least $2 - 4\delta\epsilon$. Further all outgoing vectors from $\mathcal{X}_I$ are located within $4\delta\epsilon$ distance of each other. The condition that the distance between any two states from the *same* subchain is less than the distance between any two *different* subchains is clearly satisfied when $d_{out} \geq 2 - 4\delta\epsilon > 2d_{in}$; that is, $\epsilon < \frac{1-\eta}{6\delta}$. □

This threshold provides a condition that guarantees worst-case separability, since $d_{out} > 2d_{in}$ is a conservative condition for being separable. We now characterize the probability of correct classifications, for which we make the following assumption regarding the distribution of the entries in $\mathbf{C}$.

Assumption *2:* All entries in off-diagonal blocks of the perturbation matrix $\mathbf{C}$ follow independent, identical *half-normal distributions* with standard deviation equal to 1, that is, $C_{i_I j_J} = |\hat{C}_{i_I j_J}|, \forall i, j, I \neq J$, where $\hat{C}_{i_I j_J} \sim \mathcal{N}(0, 1)$.

Theorem *2:* Let $\mathcal{N}_I$ and $\mathcal{N}_J$ be two noncommunicating subchains of a CD Markov chain $\mathcal{X}$ with transition matrix $\mathbf{P}^*$. The probability of distinguishing the I^{th} and the J^{th} subchains is given by

$$\prod_{i,i'\in I,\, j\in J}\mathbb{P}\{\|\mathbf{p}(i_I) - \mathbf{p}(i'_I)\|_2^2 < \|\mathbf{p}(j_J) - \mathbf{p}(i_I)\|_2^2\}, \tag{9}$$

which leads to lower bounds on the separability given by

$$\begin{aligned} \mathbb{P}\Big\{ &(n - n_I)\epsilon^2 Y_1 + \sqrt{2}\epsilon Y_2 + n\epsilon^2\left(\tfrac{1}{n_I} + \tfrac{n}{n_J}\right)Y_3 \\ &-\sqrt{2}\epsilon Y_4 < \tfrac{1}{n_I} + \tfrac{1}{n_J} - \eta^2 \Big\}; \end{aligned} \tag{10}$$

Y_i are Chi-squared random variables $Y_i \sim \chi^2(df_i), i = 1, 2, 3, 4$, with degree of freedom $df_1 = df_2 = 2(n - n_J), df_3 = n$ and $df_4 = n - n_I - n_J$.

The proof of Theorem 2 relies on *Assumption 2* and the resulting independence of the row vectors of $\mathbf{C}$; see [18] for details.

B. Dynamics of the aggregated Markov chain

We first discuss aggregation of NCD Markov chains as proposed by Simon and Ando [2]. In keeping with the notation we use, $\mathbf{P}$ represents a transition matrix for an NCD Markov chain with n states; that is, $\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}$ where $\mathbf{P}^* = \text{diag}\{\mathbf{P}_I^*\}_{I=1}^N$ is a CD stochastic matrix and $\mathbf{C}$ is a perturbation matrix satisfying constraints (2). Simon and Ando show that if $\boldsymbol{\pi}_I^*$ is the stationary distribution vector of $\mathbf{P}_I^*$, then the matrix $\mathbf{Q}^\circ$ with

$$Q_{IJ}^\circ = \sum_{i=1}^{n_I} \pi_{i_I}^* \sum_{j=1}^{n_J} P_{i_I j_J}, \tag{11}$$

is a stochastic matrix that defines another Markov chain with N superstates. Moreover, the stationary distribution of $\mathbf{Q}^\circ$ satisfies

$$\|\boldsymbol{\pi}(\mathbf{Q}^\circ) - \boldsymbol{\pi}(\mathbf{P})\boldsymbol{\Phi}\|_1 \sim O(\epsilon^2), \tag{12}$$

where $\boldsymbol{\Phi}$ is the aggregation matrix corresponding to $\mathbf{P}^*$. Specifically, it has been shown that every entry of the vector $\boldsymbol{\pi}(\mathbf{Q}^\circ) - \boldsymbol{\pi}(\mathbf{P})\boldsymbol{\Phi}$ is in order ϵ^2. An improved approximation error in the steady state distribution is given in [3].

In order to implement this state aggregation and compute $\mathbf{Q}^\circ$, it is necessary to know the correct state partition for the subchains, and the stationary distribution of each $\mathbf{P}_I^*$; both of these quantities are not always available. In fact, most modeling methods yield the $\mathbf{P}$ matrix, and recovering the underlying decomposable subchains is one of the objectives.

In using the DA-based algorithm to aggregate a large Markov chain, we need to specify state weights. These weights influence the aggregation resulting from the alternating minimization of the dissimilarity function $\nu(\mathbf{P}, \mathbf{Q})$ over $\boldsymbol{\Phi}$ and $\mathbf{Z}$ in (3). In the following, we discuss the influence of the initial state weight vector on the stationary distribution of the aggregated chain. The main results are given in *Theorem 3*. We first present a useful Lemma.

Lemma *1:* For an NCD Markov chain as given by (1), the following hold

(a) the transition matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$ of a correctly identified subchain structure obtained from the DA-based aggregation algorithm is given by

$$\mathbf{Q} = \mathbf{R}^T\mathbf{P}\boldsymbol{\Phi} \Leftrightarrow Q_{IJ} = \sum_{i=1}^{n_I} \frac{\mu_{i_I}}{\mu_I} \sum_{j=1}^{n_J} P_{i_I j_J},$$

where $\boldsymbol{\Phi}$ is the aggregation matrix and $\mathbf{R} = \text{diag}(\boldsymbol{\mu})\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^T\text{diag}(\boldsymbol{\mu})\boldsymbol{\Phi}\right)^{-1} \in \mathbb{R}_+^{n\times N}$, and $\mu_I = \sum_{i=1}^{n_I} \mu_{i_I}$ is the weight of the I^{th} group of states.

(b) Let $\mathbf{G}_I^*$ and $\mathbf{G}^*$ be defined in terms of the left eigenvectors of $\mathbf{P}^*$ as $\mathbf{G}_I^* = \left[\boldsymbol{\pi}^*(1_I) \quad \cdots \quad \boldsymbol{\pi}^*(n_{II})\right]^T$ and $\mathbf{G}^* = \text{diag}(\mathbf{G}_I^*)$. If $\{\lambda(l_L), \boldsymbol{\theta}(l_L)^T\}$ is the eigenvalue/left eigenvector pair for the matrix $\mathbf{G}^*\mathbf{P}(\mathbf{G}^*)^{-1}$, then the vector $\boldsymbol{\theta}_1(l_L)^T \stackrel{\triangle}{=} [\theta_{1_1}(l_L), \ldots, \theta_{1_N}(l_L)]$ satisfies $\boldsymbol{\theta}_1(l_L)^T = \boldsymbol{\pi}^T(l_L)\boldsymbol{\Phi}$. In particular, when $L = 1$ we have

$$\boldsymbol{\theta}_1(1_1)^T = \boldsymbol{\pi}(\mathbf{P})^T\boldsymbol{\Phi}, \tag{13}$$

where $\boldsymbol{\pi}(\mathbf{P})^T$ is the stationary distribution of $\mathbf{P}$.

(c) The vector $\boldsymbol{\theta}(l_L)^T$ satisfies the equation $\mathbf{A}\boldsymbol{\theta}(l_L) = \epsilon\mathbf{b}$, where

$$A_{IJ} = \begin{cases} 1 - \lambda(l_L) & \text{if } J = I \\ \epsilon n_I \sigma_{1_I 1_J} & \text{else} \end{cases}, \tag{14}$$

$b_I = -n_I \sum_{J=1}^N \sum_{j=2}^{n_J} \theta_{j_J}(l_L)\sigma_{i_I i_I, \sigma_{i_I j_J}} \stackrel{\triangle}{=} \boldsymbol{\pi}^*(i_I)^T\mathbf{C}_{IJ}\boldsymbol{\upsilon}^*(j_J)$, $\mathbf{C}_{IJ}$ is the IJ^{th} block of matrix $\mathbf{C}$ and every entry of the vector $\mathbf{b}$ is of the order ϵ.

(d) For every element of the vector $\boldsymbol{\theta}_1(1_L)^T\mathbf{Q} - \lambda(1_L)\boldsymbol{\theta}_1(1_L)^T \sim O(\epsilon)$ for all $1 \le L \le N$.

Proof: We present a sketch of the proof here. The proof for (a) relies on the fact that the transition matrix $\mathbf{Q}$ is obtained when terminating the annealing process, i.e., as $\beta \to \infty$. The formula in (13) is shown to hold by evaluating the limit $\tilde{\mathbf{Q}} \to \mathbf{Q}$ as $\beta \to \infty$ via consideration of the element-wise limits of $\tilde{\boldsymbol{\Phi}}$ from and $\tilde{\mathbf{R}}$, which are well-defined (exist and finite).

The proof for (b) uses the formulation $(\mathbf{G}_I^*)^{-1} = \left[\frac{\boldsymbol{\upsilon}^*(1_I)}{s^*(1_I)}, \ldots, \frac{\boldsymbol{\upsilon}^*(n_{II})}{s^*(n_{II})}\right]$, where $s^*(i_I) = \boldsymbol{\pi}^*(i_I)^T\boldsymbol{\upsilon}^*(i_I)$, and $\boldsymbol{\pi}^*(i_I)^T$ and $\boldsymbol{\upsilon}^*(i_I)$ respectively represent left and right eigenvectors of $\mathbf{P}_I^*$. Since $\{\lambda(l_L), \boldsymbol{\theta}(l_L)^T\}$ is the eigenvalue/left eigenvector pair for $\mathbf{G}^*\mathbf{P}(\mathbf{G}^*)^{-1}$, and $\boldsymbol{\pi}(l_L)$ is the left eigenvector of $\mathbf{P}$, we have $\boldsymbol{\theta}(l_L)^T = \boldsymbol{\pi}(l_L)^T(\mathbf{G}^*)^{-1}$. By substituting for $(\mathbf{G}^*)^{-1}$, we obtain $\boldsymbol{\theta}_1^T(l_L) \stackrel{\triangle}{=} [\theta_{1_1}(l_L), \ldots, \theta_{1_N}(l_L)] = \boldsymbol{\pi}(l_L)^T\boldsymbol{\Phi}$.The result follows by setting $l = L = 1$.

The proof for (c) directly follows from noting that

$$\begin{aligned} \lambda(l_L)\boldsymbol{\theta}(l_L)^T &= \boldsymbol{\theta}(l_L)^T\mathbf{G}^*(\mathbf{P}^* + \epsilon\mathbf{C})(\mathbf{G}^*)^{-1} \\ &= \boldsymbol{\theta}(l_L)^T\boldsymbol{\Lambda}^* + \epsilon\boldsymbol{\theta}(l_L)^T\boldsymbol{\Sigma}\,\text{diag}(\frac{1}{s^*(i_I)}), \end{aligned}$$

where $\boldsymbol{\Sigma} = [\sigma_{i_I j_J}]$, and then studying the 1_Ith element of $\lambda(l_L)\boldsymbol{\theta}(l_L)^T$, which results in

$$\begin{aligned} &[1 - \lambda(l_L)]\,\theta_{1_I}(l_L) + \epsilon n_I \sum_{J=1}^N \theta_{1_J}(l_L)\sigma_{1_J 1_I} \\ &= -\epsilon n_I \sum_{J=1}^N \sum_{j=2}^{n_J} \theta_{j_J}(l_L)\sigma_{i_J i_I}. \end{aligned} \tag{15}$$

Note that for any $j \neq 1$, $\theta_{j_J}(l_L)$ is of the order ϵ (from matrix perturbation theory, [3], [19]). Thus each term in the summation of the right hand side of (15) is of order ϵ^2.

The proof of (d) follows from computing $\sigma_{1_J 1_I} = \boldsymbol{\pi}^*(1_J)^T\mathbf{C}_{JI}\boldsymbol{\upsilon}^*(1_I)$ and substituting in (15). Then it follows that the I^{th} element of $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N]$ given by $\sum_{j=1}^N \theta_{1_I}(1_L)Q_{JI} - \lambda(l_L)\theta_{1_I}(1_L)$ $=$ $\underbrace{\theta_{1_I}(1_L)\sum_{J\neq I}\left[\sum_{i=1}^{n_I}\left[\pi_{i_I}^*(1_L) - \frac{\mu_{i_I}}{\mu_I}\right]\sum_{j=1}^{n_I} P_{i_I j_J}\right]}_{(a_I)} +$ $\underbrace{\sum_{J\neq I}\theta_{1_J}(1_L)\left[\sum_{j=1}^{n_J}\left[\pi_{j_J}^*(1_L) - \frac{\mu_{j_J}}{\mu_J}\right]\sum_{i=1}^{n_I} P_{j_J i_I}\right]}_{(b_I)}$ $\underbrace{-\epsilon n_I \sum_{J=1}^N \sum_{j=2}^{n_J} \theta_{j_J}(1_L)\sigma_{j_J i_I}}_{(c_I)}$. Here (c_I) is of the order ϵ^2 since all $\theta_{j_J}(l_L)$s are order ϵ for $j \neq 1$. Using $\|\theta_1\|_1 \le 1$, we can show that $\sum_I |a_I| \le 2\epsilon$. Similarly $\sum_I |b_I|$ can be bounded above by 2ϵ. Therefore, the l_1 norm of $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N] = \sum_{I=1}^n |a_I + b_I - c_I| \le$

$\sum_{I=1}^{n}|a_I| + \sum_{I=1}^{n}|b_I| + \sum_{I=1}^{n}|c_I| \leq 4\epsilon + O(\epsilon^2)$, which implies this l_1 norm is of the order 4ϵ. □

Remark *1:* If we select state weight $\boldsymbol{\mu}$ such that

$$\frac{\mu_{j_J}}{\mu_J} = \pi^*_{j_J}(1_1), \quad \forall j, J,$$

terms (a) and (b) are both zero and the error results solely from the term (c). This is essentially the aggregation proposed in [2], whose error has been established as ϵ^2 in [3].

For the aggregated chain $\mathbf{Q} \in \mathbb{R}^{N\times N}$, for each $L = 1, \ldots, N$ we denote the eigenvalue and left eigenvector by $\{\kappa(L), \boldsymbol{\alpha}^*(L)^T\}$. Our goal is to find a bound for vector $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q}-\kappa(L)\mathbf{I}_N]$, with $\mathbf{Q}$'s eigenvalue κ. Here we make a technical assumption on the spectrum of $\mathbf{Q}$.

Assumption *3:* Assume the eigenvalues of $\mathbf{Q}$ are distinct, specifically, assume there exists a positive number Δ, such that $\max_{1\leq K,K'\leq N}|\kappa(K) - \kappa(K')| \geq \Delta$.

Theorem *3:* For an NCD Markov chain $\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}$ as in (1), let $\boldsymbol{\pi}(\mathbf{P})$ be the stationary distribution of $\mathbf{P}$. Further assume $\mathbf{P}^*_I$ is aperiodic and irreducible for $I = 1, \ldots, N$, and $\boldsymbol{\pi}^*_I = [\pi_{1_I}, \ldots, \pi_{n_{I_I}}]$ is its unique stationary distribution. Let $\mathbf{Q}$ be the transition matrix for the aggregated Markov chain resulting from choosing uniform state weights $\mu_i = \frac{1}{n}\ \forall i$, and let $\boldsymbol{\pi}(\mathbf{Q})$ be the stationary distribution of $\mathbf{Q}$. Then

$$\|\boldsymbol{\pi}(\mathbf{Q}) - \boldsymbol{\pi}(\mathbf{P})\boldsymbol{\Phi}\|_1 \sim O(\epsilon).$$

Proof: Lemma 1(b) indicates that each element of vector $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N]$ is order ϵ. However, from Lemma 1(d), $\boldsymbol{\theta}_1(1_L)^T = \boldsymbol{\pi}^T(1_L)\boldsymbol{\Phi}$ must contain elements exceeding ϵ. Without less of generality, suppose this occurs in the first entry, $\theta_{1_1}(1_L)$. Then to guarantee $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N]$ is order ϵ element-wise, all elements in the first row of matrix $[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N]$ must be order ϵ. Thus we have $\det[\mathbf{Q} - \lambda(1_L)\mathbf{I}_N] \sim \epsilon$, which differs from the characteristic equation of $\mathbf{Q}$ by ϵ. Therefore, the roots of these two equations differ at ϵ, implying that for each L there exists an eigenvalue $\kappa(L')$ of $\mathbf{Q}$ (not necessarily ordered in a particular way) such that $|\kappa(L') - \lambda(1_L)| \sim \epsilon$ for some index L'. In summary, for each L, there exists some $\kappa(L')$ close to $\lambda(1_L)$.

Note that $\kappa(L')$ is the $(L')^{\text{th}}$ eigenvalue of $\mathbf{Q}$; let $\boldsymbol{\alpha}(L')^T$ be the corresponding left eigenvector, that is, $\kappa(L')\boldsymbol{\alpha}(L')^T = \boldsymbol{\alpha}(L')^T\mathbf{Q}$. We consider the following vector $[\boldsymbol{\theta}_1(1_L)^T - \boldsymbol{\alpha}(L')^T][\mathbf{Q} - \kappa(L')I_N]$ also given by $\boldsymbol{\theta}_1(1_L)^T[\mathbf{Q}-\lambda(1_L)\mathbf{I}_N]+[\lambda(1_L)-\kappa(L')]\boldsymbol{\theta}_1(1_L)^T\mathbf{I}_N$, which is of order ϵ, since both terms on the right hand side are order ϵ (since $\det[\mathbf{Q}-\lambda(1_L)\mathbf{I}_N] \sim \epsilon$ and $|\kappa(L')-\lambda(1_L)| \sim \epsilon$). On the other hand, the K^{th} entry of the vector on the left hand side is given by $[\boldsymbol{\theta}_1(1_L)^T - \boldsymbol{\alpha}(L')^T]\boldsymbol{\xi}_K$, where $\boldsymbol{\xi}_K$ is the K^{th} column of $[\mathbf{Q} - \kappa(L')\mathbf{I}_N]$. A necessary and sufficient condition to guarantee $[\boldsymbol{\theta}_1(1_L)^T - \boldsymbol{\alpha}(L')^T][\mathbf{Q} - \kappa(L')I_N] \sim O(\epsilon)$ for $L = 1$ is that $|\boldsymbol{\theta}_1(1_L) - \boldsymbol{\alpha}(L')|^T$ is order ϵ element-wise. Let the index $L'' = \arg\max_K \kappa(K)$, i.e., $|\kappa(L'') - \lambda(1_1)| \sim \epsilon$, then $\boldsymbol{\xi}_K$ exceeds order ϵ for all $K \neq L''$. Therefore, $[\boldsymbol{\theta}_1(1_1)^T - \boldsymbol{\alpha}(L')^T]_K \sim \epsilon$ for all $K \neq L''$. Moreover, since $\|\boldsymbol{\theta}_1(1_L)\|_1 = \|\boldsymbol{\alpha}(L')\|_1 = 1$, the difference of the L'' entry is also order ϵ. These give the necessity; sufficiency is straightforward. Rewrite $|\boldsymbol{\theta}_1(1_L) - \boldsymbol{\alpha}(L')| \sim \epsilon$ for $L = 1$, then we have $\|\boldsymbol{\pi}(\mathbf{Q}) - \boldsymbol{\pi}(\mathbf{P})\boldsymbol{\Phi}\|_1 \sim O(\epsilon)$ as needed. □

V. Numerical Examples

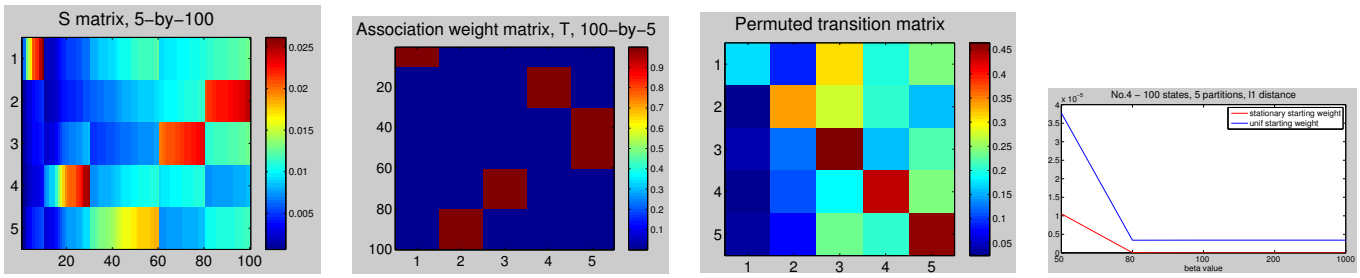
Example 1: Markov chain with index of balance $\eta = 0$: This example illustrates the analysis in *Lemma 1* when $\eta = 0$; for an NCD Markov chain $\mathbf{P} = \mathbf{P}^* + \epsilon\mathbf{C}$ satisfying *Assumption 1*, the effectiveness of the DA-based algorithm in identifying subchains is demonstrated. We design the example such that the CD component $\mathbf{P}^*$ of $\mathbf{P}$ contains repeated rows. However, after adding a perturbation term the transition matrix $\mathbf{P}$ is much less restrictive. We set $\mathbf{P}^* = \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.2 & 0.8 \end{bmatrix}$; the perturbation term is generated according to *Assumption 2* and entries are scaled so that the row sums of $\mathbf{C}$ are zero. For example, $\mathbf{C} = \begin{bmatrix} -0.4014 & -0.4325 & 0.7885 & 0.0455 \\ -0.7205 & -0.3341 & 0.5408 & 0.5139 \\ 0.4456 & 0.6599 & -0.2130 & -0.8925 \\ 0.4914 & 0.5054 & -0.5498 & -0.4469 \end{bmatrix}$. In this case the maximum degree of coupling $\delta = 1.1055$, using (5). The result of *Theorem 1* indicates $\epsilon < \frac{1}{6\delta} = 0.1508$. We choose $\epsilon = 0.15$, giving us $\mathbf{P} = \begin{bmatrix} 0.6398 & 0.2351 & 0.1183 & 0.0068 \\ 0.5919 & 0.2499 & 0.0811 & 0.0771 \\ 0.0668 & 0.0990 & 0.1681 & 0.6661 \\ 0.0737 & 0.0758 & 0.1175 & 0.7330 \end{bmatrix}$;note $\mathbf{P}$ does not have a significant "repeated row" structure. For this small example, we can easily compute the stationary distribution of $\mathbf{P}$, and select $\boldsymbol{\mu}$ by the stationary weights. The algorithm successfully aggregates the first two states and the last two states. The resulting $\mathbf{Q}$ matrix for the aggregated chain is given by $\mathbf{Q} = \begin{bmatrix} 0.6909 & 0.3091 \\ 0.2718 & 0.7282 \end{bmatrix}$. The stationary distributions of a direct aggregation of $\boldsymbol{\pi}(\mathbf{P})$ and $\boldsymbol{\pi}(\mathbf{Q})$ at different stages of the algorithm (indicated by value β) are given in the table below. It is evident that the DA-based clustering algorithm not only identifies correct subchains, but also provides increasingly better approximation to the limiting distribution of $\mathcal{X}$.

β	$\boldsymbol{\pi}(\mathbf{P})\tilde{\boldsymbol{\Phi}}$	$\boldsymbol{\pi}(\mathbf{Q})$	l_1 distance
10	[0.5300, 0.4700]	[0.5321, 0.4679]	3×10^{-3}
20	[0.5321, 0.4679]	[0.5319, 0.4681]	2.6981×10^{-4}
50	[0.5321, 0.4679]	[0.5321, 0.4679]	2.1776×10^{-8}

Example 2: Markov chain with non-repeated rows: We now consider a larger NCD Markov chain with 100 states whose transition matrix is represented by Figure 1 (a). There are 5 set of states within which transitions between states are highly possible. Each partition contains $10, 20, 30, 20, 20$ states, respectively, and the states are ordered such that the blocked diagonal structure is clearly observable.

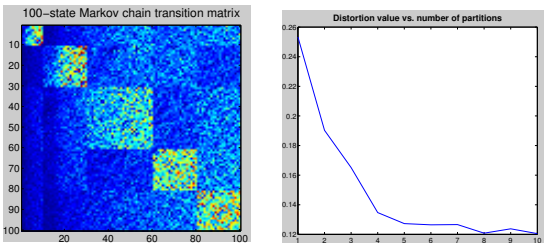
If we fix the size of $\mathbf{Q}$ to be 5, then regardless of the permutation of the states of $\mathbf{P}$, our clustering algorithm always identifies the correct subchains. The soft aggregation matrix $\tilde{\boldsymbol{\Phi}} \in [0, 1]^{100\times 5}$, the resulting composite transition matrix $\mathbf{Z} \in \mathbb{R}^{5\times 100}$, and the corresponding transition matrix for superstates $\mathbf{Q} \in \mathbb{R}^{5\times 5}$ are shown in Figure 2 (a) and (b). The $\tilde{\boldsymbol{\Phi}}$ matrix demonstrates a nearly deterministic partition and the superstate ordering. By adopting the same permutation on the $\mathbf{Q}$ matrix, we obtain a transition matrix corresponding to the 5 superstates in the original order (Figure 2 (c)).

The aggregated stationary distribution of $\mathbf{P}$ is given by



(a) (b) (c) (d)

Fig. 2. The scaled color plots of clustering results for the 100-state Markov chain given by Figure 1. Plots (a) and (b) show the composite transition matrix $\mathbf{Z}$ and the soft aggregation matrix $\tilde{\mathbf{\Phi}}$, solved using the relaxed optimization problem (4). Plot (c) is the transition matrix of the 5-state chain after applying the original sub-chain reordering. (d) demonstrates the insensitivity to the initial state weight. The l_1 distances between the limiting distribution of the aggregated chain, and the true partitions, $\|\boldsymbol{\pi}(\mathbf{Q}) - \boldsymbol{\pi}(\mathbf{P})\mathbf{\Phi}\|$ are shown.



(a) (b)

Fig. 1. (a) - Transition probability matrix $\mathbf{P}$ of the 100-state Markov chain, warm color indicates a larger value, and a colder color indicates a smaller (nonnegative) value. (b) - The aggregation errors (soft dissimilarity $\tilde{\nu}(\mathbf{P}, \mathbf{Q})$) achieved by the lower-order Markov chain $\mathbf{Q}$ resulted from clustering, as function of size of $\mathbf{Q}$.

$\boldsymbol{\pi}\mathbf{\Phi} = [0.0365\ 0.1358\ 0.2990\ 0.2468\ 0.2819]$. Note the stationary distribution of the aggregated Markov chain $\mathbf{Q}$ meets $\boldsymbol{\pi}\mathbf{\Phi}$ with high accuracy. Moreover, this is independent of the initial selection of state weights. If we repeat the procedure with two choices of state weights, (1) $\boldsymbol{\mu}_s := \boldsymbol{\pi}$ being proportional to the stationary distribution, and (2) $\boldsymbol{\mu}_u$ being uniform over all states, the l_1 distances between $\boldsymbol{\pi}(\mathbf{Q})$ and $\boldsymbol{\pi}(\mathbf{P})\mathbf{\Phi}$ are plotted in Figure 2 (d). As the annealing parameter β increases (pushing the soft aggregation matrix T towards a hard limit), both choices of initial weights $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_u$ achieve excellent accuracy in the limiting distribution (note the scale of the y-axis is 10^{-5}). Considering the ease of implementation using $\boldsymbol{\mu}_u$, this will be an excellent choice for high-order systems.

References

[1] R. Srikant, *The Mathematics of Internet Congestion Control*. Birkhauser, 2004.

[2] H. A. Simon and A. Ando, "Aggregation of variables in dynamic systems," *Econometrica: journal of the Econometric Society*, pp. 111–138, 1961.

[3] P.-J. Courtois, "Error analysis in nearly-completely decomposable stochastic systems," *Econometrica: Journal of the Econometric Society*, pp. 691–709, 1975.

[4] Y. Xu, S. Kim, S. M. Salapaka, C. L. Beck, and T. P. Coleman, "Clustering large networks of parametric dynamic generative models," in *IEEE 51st Annual Conference on Decision and Control (CDC),*. IEEE, 2012, pp. 5248–5253.

[5] C. Quinn, T. P. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J.Computational Neuroscience*, Jan 2011.

[6] R. J. Fletcher, M. Acevedo, B. E. Reichert, K. Pias, and W. M. Kitchens, "Social network models predict movement and connectivity in ecological landscapes," *PNAS*, no. 48, 2011.

[7] H. Ma and A. Zeng, "The connectivity structure, giant strong component and centrality of metabolic networks," *Bioinformatics*, pp. 1423–1430, 2003.

[8] G. Varoquaux, A. Gramfort, J. Poline, and B. Thirion, "Markov models of brain connectivity from fMRI studies: is brain functional connectivity small world, or decomposable into networks?" *Journal of Physiology*, vol. 106, 2012.

[9] Y. Xu, S. M. Salapaka, and C. L. Beck, "Reduction of graphs and Markov chain models by deterministic annealing," *IEEE Transactions on Automatic Control*, vol. to appear, 2014.

[10] K. Rose, "Deterministic annealing for clustering, compression, classification, regression and related optimization problems," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov. 1998.

[11] A. Ando and F. M. Fisher, "Near-decomposability, partition and aggregation, and the relevance of stability discussions," *International Economic Review*, vol. 4, no. 1, pp. 53–67, 1963.

[12] R. Phillips and P. Kokotovic, "A singular perturbation approach to modeling and control of Markov chains," *IEEE Transactions on Automatic Control*, vol. 26, no. 5, pp. 1087 – 1094, oct 1981.

[13] K. Deng, P. Mehta, and S. Meyn, "Optimal Kullback-Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. on Automatic Control*, vol. 56, no. 12, pp. 2793–2808, Dec. 2011.

[14] Y. Xu, S. M. Salapaka, and C. L. Beck, "A distance metric between directed weighted graphs," in *52nd IEEE Conference on Decision and Control*, Dec 2013, pp. 6359–6364.

[15] ——, "On reduction of graphs and Markov chain models," in *50th IEEE Conference on Decision and Control and European Control Conference, 2011*. IEEE, 2011, pp. 2317–2322.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[17] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[18] Y. Xu, "Clustering, coverage and aggregation methods for large networks," Ph.D. dissertation, Dept. Industrial Engineering, Univ. Illinois Urbana-Champaign, Urbana, IL, USA, Tech. Rep., 2014.

[19] J. Wilkinson, *The Algebraic Eigenvalue Problem*, ser. Monographs on numerical analysis. Clarendon Press, 1988.