Supplementary Information for Dynamics Reversibility and A New Theory of Causal Emergence

Supplementary A Theorems and Proves

A.1 Propositions and Proves for *EI*

We will layout the propositions about EI and give the proves for them in this subsection.

Lemma 1. *The EI can reach its minimum value 0 if and only if the row vectors of the probability transition matrix P are identical.*

Proof. The EI can be expressed as:

$$EI = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(P_i||\bar{P}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}\log\frac{N\cdot p_{ij}}{\sum_{k=1}^{N} p_{kj}}, \tag{A1}$$

where, P_i is the ith row vector in P, and log is the logarithm function with base 2. Without losing generality, suppose EI is differentiable on p_{ij}. By taking the derivative, and using the normalization condition of probability distribution $\sum_{j=1}^{N} p_{ij} = 1$ for $1 \le i \le N$, we obtain that:

$$\frac{\partial EI}{\partial p_{ij}} = \log\left(\frac{p_{ij}}{p_{iN}}\right) - \log\left(\frac{\bar{p}_{\cdot j}}{\bar{p}_{\cdot N}}\right) \tag{A2}$$

for $1 \le i \le N$ and $1 \le j \le N-1$, where $\bar{p}_{\cdot j} = \frac{1}{N}\sum_{k=1}^{N} p_{kj}$ for $1 \le j \le N$. Here, the placeholder $\cdot$ represents to the average the probabilities of p_{kj} for all row indices. Therefore, Equation A2 is equal to 0 if and only if:

$$p_{ij} = p^*_{ij} = \bar{p}_{\cdot j} = \frac{1}{N}\sum_{k=1}^{N} p_{kj} \tag{A3}$$

for any $1 \le i, j \le N$, where p^*_{ij} represents the optimal solution of Equation A2. That is to say, all the row vectors are identical: $P_i = P_j = \bar{P}$ for any $1 \le i, j \le N$. And the corresponding value of EI is:

$$EI_{min} = 0. \tag{A4}$$

To guarantee that EI is differentiable, we require that $p_{ij} > 0, p_{iN} > 0, \bar{p}_{\cdot j} > 0$, and $\bar{p}_{\cdot N} > 0$. □

Therefore, EI can reach its minimum value 0 when all the row vectors are identical. For the special matrix $P = \frac{1}{N}\cdot\mathbb{1}$, EI also equals 0, however, it is not the unique minimum point. Actually, all the matrix with identical normalized probability row vector can make $EI = 0$.

By further taking the second order derivative of EI, we can prove that EI is not a convex function.

Corollary 1. *The second order derivative of EI with respect to the distribution p_{st} with $1 \leq s \leq N$ and $1 \leq t \leq N-1$ is:*

$$\frac{\partial^2 EI}{\partial p_{ij} \partial p_{st}} = \frac{1}{N} \cdot \left(\frac{\delta_{i,s}\delta_{j,t}}{p_{ij}} + \frac{\delta_{i,s}}{p_{iN}} - \frac{\delta_{j,t}}{N \cdot \bar{p}_{\cdot j}} - \frac{1}{N \cdot \bar{p}_{\cdot N}} \right), \tag{A5}$$

and the $EI(\{p_{ij}\})$ is not a convex function.

Proof. By further taking the derivative of Equation A2 with respect to the distribution p_{st} with $1 \leq s \leq N$ and $1 \leq t \leq N-1$, we obtain Equation A5.

When $i = s$, the second order derivative of EI is:

$$\begin{aligned} \frac{\partial^2 EI}{\partial p_{ij} \partial p_{it}} &= \frac{\delta_{j,t}}{N} \left(\frac{1}{p_{ij}} - \frac{1}{N \cdot \bar{p}_{\cdot j}} \right) + \frac{1}{N \cdot p_{iN}} - \frac{1}{N^2 \cdot \bar{p}_{\cdot N}} \\ &= \delta_{j,t} \frac{\sum_{k=1}^{N-1} p_{kj} - p_{ij}}{N^2 \cdot p_{ij} \cdot \bar{p}_{\cdot j}} + \frac{\sum_{k=1}^{N-1} p_{kN} - p_{iN}}{N^2 \cdot p_{iN} \cdot \bar{p}_{\cdot N}} \\ &= \delta_{j,t} \frac{\sum_{k \neq i} p_{kj}}{N^2 \cdot p_{ij} \cdot \bar{p}_{\cdot j}} + \frac{\sum_{k \neq i} p_{kN}}{N^2 \cdot p_{iN} \cdot \bar{p}_{\cdot N}} \geq 0, \end{aligned} \tag{A6}$$

but when $i \neq s$,

$$\frac{\partial^2 EI}{\partial p_{ij} \partial p_{st}} = -\frac{\delta_{j,t}}{N^2 \cdot \bar{p}_{\cdot j}} - \frac{1}{N^2 \cdot \bar{p}_{\cdot N}} < 0 \tag{A7}$$

holds, no matter if $j = t$ or not. Therefore, the Hessian matrix of EI is not positive-definite, EI is not a convex function. □

We will further discuss the condition and the properties of the maximum of EI.

Lemma 2. *The EI measure can reach its maximum $\log N$ if and only if P is a permutation matrix.*

Proof. By noticing that

$$\begin{aligned} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} \log \bar{p}_{\cdot j} &= \sum_{j=1}^{N} \left(\sum_{i=1}^{N} p_{ij} \right) \log \bar{p}_{\cdot j} \\ &= \sum_{j=1}^{N} \bar{p}_{\cdot j} \log \bar{p}_{\cdot j} = -H(\bar{P}), \end{aligned} \tag{A8}$$

where $H(\bar{P})$ is the Shannon entropy of the average distribution $\bar{P} \equiv \sum_{i=1} P_i/N$. Thus, EI can also be separated as:

$$\begin{aligned} EI &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}\log p_{ij} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\bar{p}_{\cdot j}\log\bar{p}_{\cdot j} \\ &= \frac{1}{N}\sum_{i=1}^{N}(-H(P_i)) + H(\bar{P}). \end{aligned} \tag{A9}$$

Where $H(P_i) = -\sum_{j=1}^{N} p_{ij}\log p_{ij}$ is the Shannon entropy of the distribution P_i. Furthermore, we have:

$$-H(P_i) \leq 0, \tag{A10}$$

the equality holds when P_i is a one hot vector, and we also have:

$$H(\bar{P}) \leq \log N, \tag{A11}$$

and the equality holds when $\bar{P} = \frac{1}{N}\cdot\mathbb{1}$. By combining these two inequalities together, we can obtain:

$$EI \leq 0 + \log N = \log N. \tag{A12}$$

The condition that make the equality in Equation A10 and the equality in Equation A11 hold simantanously is that all row vectors in P are one-hot vectors, and they are all different such that

$$\frac{1}{N}\sum_{i} P_i = \bar{P} = \frac{1}{N}\cdot\mathbb{1}. \tag{A13}$$

Therefore, we reach the statement claimed by this theorem that P must be a permutation matrix.

□

A.2 Theorems and Proves for Dynamical Reversibility and Γ

A.2.1 Dynamical Reversibility and Time Reversibility

Proposition 1. *For a given Markov chain χ and the corresponding TPM P, with P being dynamically reversible as defined in Definition 1, if and only if P is a permutation matrix.*

Proof. If P is dynamical reversible, then P must be an invertible matrix and P is a TPM (which means all row vectors are normalized $|P_i|_1 = 1$).

Because P is invertible, therefore,

$$P = U \cdot diag(\lambda_1, \lambda_2, \cdots, \lambda_N) \cdot U^{-1}, \tag{A14}$$

where, U is an orthonormal matrix, $\lambda_1, \lambda_2, \cdots, \lambda_N$ are the eigenvalues of P, and their modulus are less than or equal to 1:

$$|\lambda_i| \leq 1, \forall i \in [1, N], \tag{A15}$$

these inequality holds because P is a TPM of a Markov chain according to [50].

Thus, the inverse of P can be expressed by:

$$P^{-1} = U \cdot diag(\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_N^{-1}) \cdot U^{-1}, \tag{A16}$$

and:

$$|\lambda_i|^{-1} \geq 1, \forall i \in [1, N], \tag{A17}$$

However, this conflicts with the conclusion that the modulus of all the eigenvalues of a TPM must be less or equals to 1 if the inequality holds strictly. Thus, we have:

$$|\lambda_1| = |\lambda_2| = \cdots = |\lambda_N| = 1. \tag{A18}$$

Therefore, the eigenvalues are the complex solutions for the equation of $x^N = 1$. Suppose the ith eigenvalue is λ_i and the corresponding eigenvector is v_i, therefore:

$$Pv_i = \lambda_i v_i \tag{A19}$$

By taking conjugate transpose on two sides of Equation A19 and multiplying back to this equation, we get:

$$v_i^\dagger P^\dagger P v_i = \lambda_i \lambda_i^* v_i^\dagger v_i = v_i^\dagger v_i. \tag{A20}$$

where † is the conjugate transpose operator, * stands for the conjugate complex number.

This equation holds for any i, and consider P is positive real matrix, we conclude:

$$P^T \cdot P = P \cdot P^T = I \tag{A21}$$

Therefore, any two row vectors of P must satisfy:

$$P_i \cdot P_j = \delta_{ij}, \forall i, j \in \{1, 2, \cdots, N\} \tag{A22}$$

where $\cdot$ represents the scalar product between P_i and P_j, $\delta_{ij} = 1$ only if $i = j$, otherwise $\delta_{ij} = 0$. While, because $1 \geq p_{ik} \geq 0, \forall i, k \in \{1, 2, \cdots, N\}$, thus P_is must be one-hot vectors, and P_i is orthogonal to P_j for any $i, j \in \{1, 2, \cdots, N\}$. Therefore, P is a permutation matrix.

On the other hand, if P is a permutation matrix, then it is invertible because permutation matrices are full rank and eigenvalues are 1. The normalization conditions $|P_i|_1 = 1, \forall i \in \{1, 2, \cdots, N\}$ are satisfied because all row vectors P_i are one-hot vectors. Therefore, P is dynamical reversible. □

Next, we will prove dynamical reversibility implies time reversibility for Markov chains.

Lemma 3. *For a recurrent discrete Markov chain χ, suppose its TPM is P on the space of states $\mathcal{S}$ and its stationary distribution is μ, if P is dynamically reversible, then χ also satisfies time reversibility.*

Proof. If P is dynamically reversible, then P is a permutation matrix according to Proposition 1, thus, the corresponding stationary distribution must be $\mathbb{1}/N$, where

$\mathbb{1} = (1, 1, \cdots, 1)$, such that any permutation on elements in $\mathbb{1}/N$ is the same vector. Because P is invertible, it satisfies

$$P = P^{-1} = P^T, \tag{A23}$$

then we have

$$p_{ij} \cdot \frac{1}{N} = p_{ji} \cdot \frac{1}{N}. \tag{A24}$$

If we denote $\mu = \frac{\mathbb{1}}{N}$ as a row vector, Equation A24 can be written in the following form:

$$\mu_i \cdot p_{ij} = \mu_j \cdot p_{ji}, \forall i, j \in \{1, 2, \cdots, N\} \tag{A25}$$

This is the condition for the time reversible Markov chain[32]. Therefore, χ is time reversible, and P^T is the TPM of the reversible process.

□

A.2.2 Approximate Dynamical Reversibility Γ_α

We will present the propositions and theorems and the proofs about the measure of approximate dynamical reversibility Γ_α in this sub-section. Before the propositions are presented, we will prove the lemma related with the proposition that will be used in the following parts.

Lemma 4. *For a TPM $P = (P_1^T, P_2^T, \cdots, P_N^T)^T$, where P_i is the i-th row vector, then:*

$$P_i \cdot P_j \leq 1, \forall i, j \in \{1, 2, \cdots, N\}, \tag{A26}$$

where, $\cdot$ represents the scalar product between the vectors P_i and P_j. The equality holds when $P_i = P_j$ and is a one-hot vector.

Proof. Because P_i is a probability distribution, therefore, it should satisfy normalization condition which can be expressed as:

$$|P_i|_1 = \sum_{j=1}^{N} p_{ij} = 1, \tag{A27}$$

where, $|\cdot|_1$ is 1-norm for vector, which is defined as the summation of the absolute values of all elements. Because P_i and P_j are all positive vectors, we have:

$$P_i \cdot P_j = \sum_{k=1}^{N} p_{ik} p_{jk} \leq \sum_{k=1}^{N} \sum_{l=1}^{N} p_{ik} p_{jl} = |P_i|_1 \cdot |P_j|_1 = 1. \tag{A28}$$

for any $i, j \in \{1, 2, \cdots, N\}$. The equality holds when

$$\sum_{k \neq j} p_{ik} p_{jl} = 0. \tag{A29}$$

Due to $p_{ij} > 0, \forall i, j \in \{1, 2, \cdots, N\}$, Equation A29 holds only if $P_i = P_j$ and P_i is a one-hot vector. □

Lemma 5. *For a given TPM P, suppose its singular values are $(\sigma_1, \sigma_2, \cdots, \sigma_N)$, we have:*

$$\sum_{i=1}^{N} \sigma_i^2 = \sum_{i=1}^{N} P_i^2 = ||P||_F^2 \leq N, \tag{A30}$$

and the equality holds when $P_i, \forall i \in \{1, 2, \cdots, N\}$ are one-hot vectors.

Proof. Because $\sigma_i^2, \forall i \in \{1, 2, \cdots, N\}$ are the eigenvalues of $P \cdot P^T$, thus, P can be written:

$$P \cdot P^T = U \Sigma^2 U^T, \tag{A31}$$

where U is a orthonormal matrix with size N, $\Sigma^2 = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2)$. Thus,

$$\begin{aligned} \sum_{i=1}^{N} \sigma_i^2 = \operatorname{Tr} \Sigma^2 = \operatorname{Tr} \left(U \Sigma^2 U^T\right) = \operatorname{Tr} \left(P \cdot P^T\right) \\ = \sum_{i=1}^{N} P_i^2 \leq N. \end{aligned} \tag{A32}$$

The last inequality holds because of Lemma 4. And the equality holds if and only if:

$$P_i \cdot P_i = 1, \forall i \in \{1, 2, \cdots, N\}, \tag{A33}$$

If $P_i \cdot P_i = 1$, then $\sum_j p_{ij}^2 = 1$, indicating that P_i is on a unit ball. Additionally, $|P_i|_1 = \sum_{j=1}^{N} p_{ij} = 1$ for all i, implying that P_i is on the unit hyperplane. Given that $0 \leq p_{ij} \leq 1$ for all $i, j \in \{1, 2, \cdots, N\}$, the intersection of the unit sphere and the unit hyperplane occurs at the corners. Consequently, P_i must be a one-hot vector, where only one element is 1 and the rest are zeros.

Further, it should be noticed that:

$$\sum_{i=1}^{N} \sigma_i^2 = \sum_{i=1}^{N} P_i^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^2 = ||P||_F^2 \tag{A34}$$

Therefore, Equation A30 holds, and the equality in Equation A32 holds when all row vectors are one-hot vectors. □

Lemma 6. *For a TPM P, we can write it in the following way:*

$$P = (P_1^T, P_2^T, \cdots, P_N^T)^T, \tag{A35}$$

where P_i is the i-th row vector. And suppose P's singular values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$. Thus, if

$$P_i \cdot P_i = 1, \forall i \in \{1, 2, \cdots, N\}, \tag{A36}$$

then the singular values of P satisfy:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 1 \tag{A37}$$

and

$$\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_N = 0, \tag{A38}$$

where r is the rank of the matrix P.

Proof. If P is formed by one-hot vectors, and the rank of P is r, then there are r different row vectors. Suppose they are $e_{i_1}, e_{i_2}, \ldots, e_{i_r}$, and they repeat $n_1, \ldots, n_r$ times,respectively, where $n_1 + n_2 + \cdots + n_r = N$.

Equivalently, there are r nonzero columns and there are n_1 ones in the i_1th column, n_2 ones in the i_2th column, $\ldots$, and n_r ones in the i_rth column. Those ones lie in different rows. So $P^T \cdot P$ is a diagonal matrix with $n_1, n_2, \ldots, n_r$ and zeros as its diagonal elements. So the nonzero singular values of P are $\sigma_1 = \sqrt{n_1}, \sigma_2 = \sqrt{n_r}, \ldots, \sigma_r = \sqrt{n_r}$. Since n_i is positive integer, and we can assume that $n_1 \geq n_2 \geq \cdots \geq n_r$, so

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 1$$

and

$$\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_N = 0,$$

□

We want to prove that it is reasonable that the proposed measure Γ_α to characterize the dynamical reversibility. First, we will prove that Γ_α is upper bounded by the system size N, and it can reach the maximum value if and only if P is reversible.

Lemma 7. *For a given TPM P, the measure of dynamical reversibility Γ_α for any $\alpha \in (0, 2)$,is less than or equal to the size of the system N.*

Proof. Because $0 < \alpha < 2$, $f(x) = x^{\alpha/2}$ is a concave function, and according to Lemma 5, we have:

$$\begin{aligned} \Gamma_\alpha = \sum_{i=1}^{N} \sigma_i^\alpha = N \cdot \frac{\sum_{i=1}^{N} \sigma_i^{2\cdot\frac{\alpha}{2}}}{N} &\leq N \left(\frac{\sum_{i=1}^{N} \sigma_i^2}{N} \right)^{\alpha/2} \\ = N^{1-\alpha/2} \left(\sum_{i=1}^{N} P_i^2 \right)^{\alpha/2} &\leq N^{1-\frac{\alpha}{2}+\frac{\alpha}{2}} = N \end{aligned} \tag{A39}$$

□

Next, we will discuss about the upper bound of Γ_α.

Proposition 2. *The maximum of Γ_α is N for any $\alpha \in (0, 2)$, and it can be achieved if and only if P is a permutation matrix.*

Proof. First, we will prove that when the maximum N is achieved for Γ_α, P must be a permutation matrix.

Notice that the condition for P_i to satisfy the equality in the second inequality of Equation A39 is $P_i^2 = 1$ for all $i \in \{1, 2, \cdots, N\}$.

Therefore, according to Lemma 6, there are two cases need to be discussed separately.

Case 1: If there are two rows of P are identical: $P_i \cdot P_j = 1$ for some i, j, then the rank of P is $r < N$. Then, according to Lemma 6 the first r singular values satisfy:

$$\sum_{i=1}^{r} \sigma_i^2 = N, \tag{A40}$$

and $\sigma_i \geq 1$ for any $i \in \{1, 2, \cdots, r\}$.

Suppose there are $s \leq r$ singular values strictly larger than 1, then $\sigma_i^2 > \sigma_i^\alpha$ for those $1 \leq i \leq s$, thus:

$$\sum_{i=1}^{s} \sigma_i^\alpha < \sum_{i=1}^{s} \sigma_i^2, \tag{A41}$$

And for $i > s$, the singular values are either 0 or 1, and therefore $\sigma_i = \sigma_i^2$. Finally, we have:

$$\Gamma_\alpha = \sum_{i=1}^{N} \sigma_i^\alpha < \sum_{i=1}^{N} \sigma_i^2 = N, \tag{A42}$$

Thus, in this case, the equality in Equation A39 does not hold.

Case 2: If $P_i \cdot P_j \neq 1$ for any $i \neq j$, and $P_i \cdot P_i = 1$ for $\forall i \in \{1, 2, \cdots, N\}$, then according to Lemma 6, if and only if all the singular values are 1, that is,

$$\sigma_i = \sigma_j = 1, \forall i, j \in \{1, 2, \cdots, N\}, \tag{A43}$$

which implies P is invertible. Notice that, this is exactly the same condition to make the equality holds for the first inequality in Equation A39.

Second, we will prove the necessity for the maximum of Γ_α. If P is a permutation matrix, the eigenvalues of P are the singular values because P is symmetric. And the singular values satisfy:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_N = 1, \tag{A44}$$

thus,

$$\Gamma_\alpha = \sum_{i=1}^{N} \sigma_i^\alpha = N \tag{A45}$$

which achieves the maximum value of Γ_α.

□

This theorem implies $\sum_i \sigma_i^\alpha$ is a better indicator for approximate dynamical reversibility than $\sum_i \sigma_i^2$ for $\alpha < 2$ although both of them can achieve the maximized value N when P is reversible. However, if P is degenerative, $\sum_i \sigma_i^2$ is also N, but $\sum_i \sigma_i^\alpha$ is not.

Next, we will discuss the minimum value and minimum point of Γ_α.

Lemma 8. *The rank of a nonzero TPM P can achieve the minimum value of 1 if and only if the row vectors of P_i for any i are identical. In this case,*

$$\Gamma_\alpha = |P_1|^\alpha \cdot N^{\alpha/2} \tag{A46}$$

Proof. If the rank of P is 1, all the $N-1$ row vectors of $P_i, \forall i \in \{1, 2, \cdots, N\}$ can be expressed by a linear function of the first row vector P_1, thus

$$P_i = k \cdot P_1, \tag{A47}$$

with $k > 0$. However, because $|P_i|_1 = 1$, thus k must be one. Therefore:

$$P_i = P_j, \forall i, j \in \{1, 2, \cdots, N\}. \tag{A48}$$

On the other hand, if Equation A48 holds, the rank of P should be 1.
In this case,

$$P \cdot P^T = |P_1|^2 \cdot \mathbb{1}_{N\times N}, \tag{A49}$$

where $|\cdot|$ is the modulus for $\cdot$. Therefore, the eigenvalues of $P \cdot P^T$ are $(|P_1|^2 \cdot N, 0, \cdots, 0)$. Thus, the singular values of P should be $(|P_1| \cdot \sqrt{N}, 0, \cdots, 0)$, this leads to Equation A46 □

Lemma 9. *For a given TPM $P = (P_1, P_2, \cdots, P_N)^T$, $\Gamma_\alpha = \sum_{i=1}^N \sigma_i^\alpha$ can reach its minimum 1 if and only if $P_i = \frac{1}{N}(1, 1, \cdots, 1)$ for $\forall i \in \{1, 2, \cdots, N\}$.*

Proof. When $P_i = \frac{1}{N}(1, 1, \cdots, 1)$ for $\forall i \in \{1, 2, \cdots, N\}$, $|P_1| = N^{-1/2}$, and according to Lemma 8,

$$\Gamma_\alpha = \sum_{i=1}^N \sigma_i^\alpha = |P_1|^\alpha \cdot N^{\alpha/2} = N^{-\alpha/2} \cdot N^{\alpha/2} = 1 \tag{A50}$$

for any $0 \leq \alpha < 2$.

On the other hand, because the minimum value of σ_i is zero, and $\Gamma_\alpha = \sum_i \sigma_i^\alpha \geq 0$, thus Γ_α can be minimized if the number of zero singular values is maximized. Notice that the number of non-zero singular values of P is the same as the rank of P. Thus, the minimum of Γ_α can be reached when the minimized rank of P is reached. In such case, according to Lemma 8, $\Gamma_\alpha = |P_1|^\alpha \cdot N^{\alpha/2}$, thus the minimized value of Γ_α is solely dependent on $|P_1|$. While, because P_1 is a probability distribution which satisfies $|P_i|_1 = 1$, thus, $P_1 \cdot P_1$ can be minimized when all the elements are equal. Thus,

$$P_1 = \frac{1}{N} \cdot \mathbb{1}. \tag{A51}$$

Therefore, $\frac{1}{N} \cdot \mathbb{1}$ is the minimum point.

□

Next, to illustrate why the dynamics reversibility measure Γ_α increases as the probability matrix P asymptotically converges to a permutation matrix, or dynamically reversible one, we have the following lemmas and the theorem. It is worth noting that

the lemma and the theorem are well-established results in linear algebra. However, we provide our own proof for the sake of completeness and convenience.

Lemma 10. *The function* $f(\alpha) = \left(\sum_{i=1}^{N} x_i^{\alpha}\right)^{1/\alpha}$ *is a monotonic decreasing function of* α *for any* $x_i \geq 0, \forall i \in \{1, 2, \cdots, N\}$ *and* $\alpha > 0$.

Proof. Because:

$$\sum_{i=1}^{N} (x_i^{\alpha})^2 \leq \left(\sum_{i=1}^{N} x_i^{\alpha}\right)^2, \tag{A52}$$

thus:

$$\log \frac{\sum_{i=1}^{N} x_i^{2\alpha}}{\sum_{i=1}^{N} x_i^{\alpha}} \leq \log \sum_{i=1}^{N} x_i^{\alpha}. \tag{A53}$$

Further, because log is a concave function, therefore:

$$\sum_{i=1}^{N} \left(\frac{x_i^{\alpha}}{\sum_{j=1}^{N} x_j^{\alpha}}\right) \cdot \log x_i^{\alpha} \leq \log \sum_{i=1}^{N} \left(\frac{x_i^{\alpha}}{\sum_{j=1}^{N} x_j^{\alpha}} \cdot x_i^{\alpha}\right), \tag{A54}$$

thus, combining Eq. A53 and A54, we have:

$$\frac{\sum_{i=1}^{N} x_i^{\alpha} \cdot \log x_i^{\alpha}}{\sum_{i=1}^{N} x_i^{\alpha}} \leq \log \sum_{i=1}^{N} x_i^{\alpha}. \tag{A55}$$

The equality holds when $x_i = 0, \forall i \in \{1, 2, \cdots, N\}$. Notice that the right hand term is $-\left(\frac{1}{\alpha}\right)' \log \sum_{i=1}^{N} x_i^{\alpha}$, where $'$ represents the derivative with respect to α, and the left hand term is $\frac{1}{\alpha} \cdot \left(\log \sum_{i=1}^{N} x_i^{\alpha}\right)'$, thus Equation A55 implies:

$$\left(\frac{1}{\alpha} \cdot \log \sum_{i=1}^{N} x_i^{\alpha}\right)' = \frac{1}{\alpha} \cdot \left(\log \sum_{i=1}^{N} x_i^{\alpha}\right)' + \left(\frac{1}{\alpha}\right)' \log \sum_{i=1}^{N} x_i^{\alpha} \leq 0. \tag{A56}$$

Therefore, $\log f(\alpha) = \frac{1}{\alpha} \cdot \left(\log \sum_{i=1}^{N} x_i^{\alpha}\right)$ is a monotonic decreasing function of α, and so does $f(\alpha) = \left(\sum_{i=1}^{N} x_i^{\alpha}\right)^{1/\alpha}$. □

Lemma 11. *The approximate dynamical reversibility measure* Γ_{α} *is lower bounded by* $||P||_F^{\alpha}$.

Proof. Because $0 < \alpha < 2$ and $\sigma_i \geq 0, \forall i \in \{1, 2, \cdots, N\}$ but the equality can not hold for all $i \in \{1, 2, \cdots, N\}$, this means $f(\alpha) = \left(\sum_{i=1}^{N} \sigma_i^{\alpha}\right)^{1/\alpha}$ is a strictly monotonic decreasing function of α according to Lemma 10, thus:

$$\Gamma_{\alpha}^{1/\alpha} = \left(\sum_{i=1}^{N} \sigma_i^{\alpha}\right)^{1/\alpha} \geq \left(\sum_{i=1}^{N} \sigma_i^2\right)^{1/2}, \tag{A57}$$

the equality holds only when $\sigma_i = 1$ or 0 for all $i \in [1, N]$. According to Lemma 5:

$$\left(\sum_{i=1}^{N} \sigma_i^2\right)^{1/2} = \left(\sum_{i=1}^{N} P_i^2\right)^{1/2} = \left[\mathrm{Tr}(P \cdot P^T)\right]^{1/2}. \tag{A58}$$

While,

$$\sqrt{\sum_{i=1}^{N} P_i^2} = \sqrt{\sum_{i}^{N}\sum_{j}^{N} p_{ij}^2} = ||P||_F \tag{A59}$$

Therefore:

$$\Gamma_\alpha \geq \left[\mathrm{Tr}(P \cdot P^T)\right]^{\alpha/2} = ||P||_F^\alpha. \tag{A60}$$

□

As $||P||_F$ can achieve its maximum when P_i is a one-hot vector, the lower bound of Γ_α increases as P approaches a matrix with one-hot row vectors. This can be summarized as a following proposition:

Lemma 12. *The lower bound of the approximate dynamical reversibility measure Γ_α increases with the number of one hot vectors in the TPM P.*

Proof. Because $|P_i|_1 = 1$ for all $1 \leq i \leq N$, therefore $||P_i||^2 \leq 1$ and the equality holds if and only if P_i is a one-hot vector. Further, according to Lemma 11, we have:

$$\Gamma_\alpha \geq \left[\mathrm{Tr}\left(P \cdot P^T\right)\right]^{\alpha/2} = \left(\sum_{i=1}^{N} P_i^2\right)^{\alpha/2}. \tag{A61}$$

Thus, the lower bound $\left(\sum_{i=1}^{N} P_i^2\right)^{\alpha/2}$ increased as each row vector $P_i, \forall i \in \{1, 2, \cdots, N\}$ converges to a one-hot vector. □

However, P may not be a permutation matrix because some row vectors may be similar, in which case P_i for all $i \in [1, N]$ are not orthogonal to each other but collapse to one direction. Thus, we need to further prove a proposition to exclude this case.

Lemma 13. *If P is formed by one-hot vectors, and the rank of P is r, so there are r different row vectors, and each of them repeat $n_1, n_2, \ldots, n_r$ times $(n_1 + \cdots + n_r = N)$, respectively, then $\Gamma_\alpha = \sqrt{n_1}^\alpha + \sqrt{n_2}^\alpha + \ldots \sqrt{n_r}^\alpha$. This value can reach its upper bound $r\sqrt{\frac{N}{r}}^\alpha$ when $n_1 = n_2 = \cdots = n_r = N/r$. This upper bound increases with r if N is fixed.*

Proof. If P is formed by one-hot vectors, and the rank of P is r, then there are r different row vectors. Suppose they are $e_{i_1}, e_{i_2}, \ldots, e_{i_r}$, and they repeat $n_1, \ldots, n_r$ times,respectively, where $n_1 + n_2 + \cdots + n_r = N$.

Equivalently, there are r nonzero columns and there are n_1 ones in the i_1th column, n_2 ones in the i_2th column, ..., and n_r ones in the i_rth column. Those ones lie in

different rows. So $P^T \cdot P$ is a diagonal matrix with $n_1, n_2, \ldots, n_r$ and zeros as its diagonal elements. So the nonzero singular values of P are $\sqrt{n_1}, \sqrt{n_r}, \ldots, \sqrt{n_r}$.

$$\Gamma_\alpha = \sqrt{n_1}^\alpha + \sqrt{n_2}^\alpha + \ldots \sqrt{n_r}^\alpha$$

For example,

$$P = \begin{pmatrix} 1&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0 \end{pmatrix}$$

has 6 different row vectors $e_1, e_2, e_3, e_4, e_5, e_6$. They repeat $1, 1, 1, 2, 2, 3$ times, respectively. $P^T \cdot P = diag(1, 1, 1, 2, 2, 3)$. The nonzero singular values of P are $1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{3}$.

The upper bound of $\Gamma_\alpha = \sqrt{n_1}^\alpha + \sqrt{n_2}^\alpha + \ldots \sqrt{n_r}^\alpha$ is $r\sqrt{\frac{N}{r}}^\alpha$ and it can be achieved if N is a multiple of r and $n_1 = n_2 = \cdots = n_r = N/r$. This is beacues $f(x) = x^{\frac{\alpha}{2}}$ is a concave function ($\alpha \in (0, 2)$).

$$\frac{1}{r}\sum_{i=1}^{r} f(n_i) \leq f(\frac{1}{r}\sum_{i=1}^{r} n_i) = f(N/r)$$

So $\Gamma_\alpha = \sum_{i=1}^{r} f(n_i) \leq r\sqrt{\frac{N}{r}}^\alpha$. The equality holds if and only if $n_1 = n_2 = \cdots = n_r = N/r$. For $\alpha < 2$ and a fixed N, the upper bound $r\sqrt{\frac{N}{r}}^\alpha = r^{1-\frac{\alpha}{2}} N^{\frac{\alpha}{2}}$ increases with r. □

A.3 Comparison between $\log \Gamma_\alpha$ and EI

In this subsection, we will establish the relationship between $\log \Gamma_\alpha$ and EI. First, we can synthesize the theoretical results about the minimum and maximum for both Γ_α and EI to derive the following theorem:

Proposition 3: *For any TPM P and $\alpha \in (0, 1)$, both the logarithm of Γ_α and EI share identical minimum value of 0 and one common minimum point at $P = \frac{1}{N}\mathbb{1}_{N\times N}$. They also exhibit the same maximum value of $\log N$ with maximum points corresponding to P being a permutation matrix.*

Proof. According to Lemma 1, EI has the minimum 0 when P has identical row vectors, and $P = \frac{1}{N} \cdot \mathbb{1}_{N\times N}$ satisfies this condition. While, according to Lemma 9,

$\log \Gamma_\alpha$ has also the minimum value 0 when $P = \frac{1}{N} \cdot \mathbb{1}_{N\times N}$. Therefore, EI and $\log \Gamma_\alpha$ share the same minimum.

On the other hand, according to Lemma 2, EI has the same maximum at $\log N$ when P is a permutation matrix. So does $\log \Gamma_\alpha, \forall \alpha \in (0, 2)$ according to Proposition 2.

Therefore, EI and $\log \Gamma_\alpha$ share the same minimum and maximum for any $\alpha \in (0, 2)$. □

We will prove a theorem that EI is upper bounded by $\frac{2}{\alpha} \log \Gamma_\alpha$.

Theorem 1. *For any TPM P, its effective information EI is upper bounded by $\frac{2}{\alpha} \log \Gamma_\alpha$, and lower bounded by $\log \Gamma_\alpha - \log N$.*

Proof. First, because the upper bound of the average distribution $\bar{P}$ is:

$$H(\bar{P}) \leq \log N, \tag{A62}$$

the equality holds when $\bar{P}$ is $\frac{1}{N} \cdot \mathbb{1}$. Second, according to the concavity of log function, we have:

$$\begin{aligned} -\frac{1}{N} \sum_{i=1}^{N} H(P_i) &= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} \log p_{ij} \\ &\leq \frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{j=1}^{N} p_{ij}^2 \right) \\ &\leq \log \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^2 - \log N, \end{aligned} \tag{A63}$$

and the equality holds when $p_{ij} = 1/N$ for all $1 \leq i, j \leq N$. Thus, we bring the two inequalities (Equation A62 and Equation A63) into Equation A9, we obtain:

$$EI = -\frac{1}{N} \sum_{i=1}^{N} H(P_i) + H(\bar{P}) \leq \log \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^2. \tag{A64}$$

This is the upper bound of EI and the equality holds when $P = \frac{1}{N} \cdot \mathbb{1}_{N\times N}$.

On the other hand, according to Lemma 9,

$$\log \Gamma_\alpha = \log \sum_{i=1}^{N} \sigma_i^\alpha \geq \log ||P||_F^\alpha = \frac{\alpha}{2} \log \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^2, \tag{A65}$$

the equality holds when $\sigma_i = 0, 1, \forall i \in 1, 2, \cdots, N$. Therefore:

$$EI \leq \frac{2}{\alpha} \log \Gamma_\alpha. \tag{A66}$$

Furthermore, because $EI \geq 0$ and $\Gamma_\alpha \leq N$, thus:

$$EI - \log \Gamma_\alpha \geq 0 - \log N = -\log N \tag{A67}$$

Therefore,

$$EI \geq \log \Gamma_\alpha - \log N, \tag{A68}$$

Thus, EI is lower bounded by $\log \Gamma_\alpha - \log N$. □

Tighter bounds are expected to be found in future works.

A.3.1 Quantification for Causal Emergence

Proposition 4. *For any given TPM P with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$ and rank r, and for any given $\epsilon \in [0, \sigma_1]$, the degree of causal emergence of P is:*

$$\Delta\Gamma_\alpha = \frac{\sum_{i=1}^{r_\epsilon} \sigma_i^\alpha}{r_\epsilon} - \frac{\sum_{i=1}^{N} \sigma_i^\alpha}{N}, \tag{A69}$$

and this degree satisfies the following inequality:

$$0 \leq \Delta\Gamma_\alpha \leq N - 1, \tag{A70}$$

where $r_\epsilon = \max\{i | \sigma_i > \epsilon\}$.

Proof. When $\epsilon \in (\sigma_N, \sigma_1]$, there is an integer $i \in (1, N]$ such that $\sigma_i > \epsilon$, and r_ϵ is the maximum one to satisfy this condition. Therefore:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{r_\epsilon} > \epsilon \geq \sigma_{r_\epsilon+1} \geq \cdots \geq \sigma_N \geq 0, \tag{A71}$$

thus:

$$\sum_{i=1}^{r_\epsilon} \sigma_i^\alpha > r_\epsilon \cdot \epsilon^\alpha, \tag{A72}$$

and:

$$-\sum_{i=r_\epsilon+1}^{N} \sigma_i^\alpha \geq -(N - r_\epsilon) \cdot \epsilon^\alpha. \tag{A73}$$

Therefore:

$$\frac{\sum_{i=1}^{r_\epsilon} \sigma_i^\alpha}{r_\epsilon} - \frac{\sum_{i=1}^{N} \sigma_i^\alpha}{N} = \sum_{i=1}^{r_\epsilon} \sigma_i^\alpha \left(\frac{1}{r_\epsilon} - \frac{1}{N}\right) - \frac{\sum_{i=r_\epsilon+1}^{N} \sigma_i^\alpha}{N} > r_\epsilon \epsilon^\alpha \left(\frac{1}{r_\epsilon} - \frac{1}{N}\right) - \frac{(N - r_\epsilon) \cdot \epsilon^\alpha}{N} = 0. \tag{A74}$$

While, when $\epsilon < \sigma_N$, $r_\epsilon = N$, so $\Delta\Gamma_\alpha = 0$, therefore:

$$\Delta\Gamma_\alpha \geq 0 \tag{A75}$$

Further, when $\epsilon = \sigma_N$, $r_\epsilon = N$, thus $\Delta\Gamma_\alpha = 0$.
On the other hand, because $r_\epsilon \leq N$, and according to Proposition 2, thus:

$$\sum_{i=1}^{r_\epsilon} \sigma_i^\alpha \leq \sum_{i=1}^{N} \sigma_i^\alpha \leq N. \tag{A76}$$

Since $1 \leq r_\epsilon \leq N$, therefore

$$0 \leq \frac{1}{r_\epsilon} - \frac{1}{N} \leq 1 - \frac{1}{N}. \tag{A77}$$

Thus,

$$\frac{\sum_{i=1}^{r_\epsilon} \sigma_i^\alpha}{r_\epsilon} - \frac{\sum_{i=1}^{N} \sigma_i^\alpha}{N} \leq \sum_{i=1}^{N} \sigma_i^\alpha (\frac{1}{r_\epsilon} - \frac{1}{N}) \leq N - 1. \tag{A78}$$

Combining Equation A74 and A78, we obtain Equation A70 □

Corollary 2. *For any given TPM P with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$ and rank r, according to Definition 4, causal emergence occurs if and only if $\Delta\Gamma_\alpha(\epsilon) > 0$ for some $\epsilon \geq 0$.*

Proof. **Case I**: When $\epsilon > 0$, according to Definition 4, if there exists an integer $i \in \{1, 2, \cdots, N\}$ such that $\sigma_i > \epsilon$ for any $\epsilon \in [0, \sigma_1]$, that is $\epsilon > \sigma_N$, then the vague CE occurs. Then, according to Proposition 4, $\Delta\Gamma_\alpha(\epsilon) > 0$ in this case.

Otherwise, if $\Delta\Gamma_\alpha(\epsilon) > 0$, then $r_\epsilon < N$, where $r_\epsilon = \max\{i|\sigma_i > \epsilon\}$, and therefore $\sigma_{r_\epsilon} > \epsilon$. According to Definition 4, vague CE occurs.

Case II: When $\epsilon = 0$, according to Definition 3, the rank r of P is less than N and $\sigma_j = 0, \forall j \in \{r+1, r+2, \cdots, N\}$, therefore

$$\frac{\sum_{i=1}^{r} \sigma_i^\alpha}{r} > \frac{\sum_{i=1}^{N} \sigma_i^\alpha}{N}, \tag{A79}$$

so $\Delta\Gamma_\alpha > 0$.

Otherwise, if $\Delta\Gamma_\alpha > 0$, there exists $r_{\epsilon=0} < N$ such that $r_{\epsilon=0} = \max\{i|\sigma_i > 0\}$. Thus, $\sigma_i = 0, \forall i > r_\epsilon$, that is $r_{\epsilon=0}$ is the rank of the matrix P. Therefore, clear CE occurs according to Definition 3.

□

Supplementary B The Relationship between SVD and Max EI

In this paper, we define clear and vague causal emergence using the spectrum of singular values, as outlined in Definitions 3 and 4. The rationale behind these definitions is that, approximately, a necessary condition for maximizing Effective Information (EI) is that the optimal coarse-graining strategy should allocate more probability mass to the directions of singular vectors corresponding to the largest singular values. However, this condition is not exact, as the actual coarse-graining strategy must meet

additional requirements—for instance, the coarse-grained TPM for macro-dynamics must satisfy the normalization condition for each row vector, and the grouping of micro-states must be well-defined and deterministic.

In this section, we theoretically demonstrate why the necessary condition should be satisfied approximately by examining one of the simplest case, the coarse-graining strategies can be represented by a clustering matrix and the coarse-grained TPM can be calculated by the multiplication of the matrices. Subsequently, we provide two examples to illustrate when will consistent conclusion about CE can be derived.

B.1 Theoretical Analysis for the Necessary Condition of EI Maximization

In the framework of Erik Hoel's theory of causal emergence, the occurence of causal emergence for a given Markovian dynamical systems relies on the coarse-graining strategy which maximizing the EI of the macro-dynamics after the coarse-graining.

We define a coarse-graining method using an $N \times r$ clustering matrix $\Phi = (\Phi_1^T, \Phi_2^T, \cdots, \Phi_r^T)$, where each vector Φ_i indicates the membership in the ith cluster. Specifically, $\Phi_{j,i} = 1$ if the jth micro-state belongs to the ith macro-state.

Since each microstate corresponds to a unique macrostate, all vectors Φ_i for $i \in \{1, 2, \cdots, r\}$ are mutually orthogonal. Thus, the transition probability matrix (TPM) for the coarse-grained macro-dynamics can be expressed as follows:

$$P' = D \cdot \Phi^T \cdot P \cdot \Phi. \tag{B80}$$

where $D = diag(1/\sum_{j=1}^{N} \Phi_{1,j}, 1/\sum_{j=1}^{N} \Phi_{2,j}, \cdots, 1/\sum_{j=1}^{N} \Phi_{r,j})$ is the normalization coefficients such that P' are composed by normalized row probability vectors. Equation B80 illustrates a naive coarse-graining method that collapses micro-states into macro-states by summing the collapsed probabilities across different columns and averaging across all rows in P.

While the expression is exact, its asymmetric form limits further theoretical analysis. To enable this, we normalize the clustering vectors Φ_i by dividing each by its norm, resulting in $\phi_i = \Phi_i/|\Phi_i|$ and $\phi = (\phi_1^T, \phi_2^T, \cdots, \phi_r^T)$. This allows us to approximately transform Equation B80 into a symmetric form:

$$P' \approx \phi^T \cdot P \cdot \phi, \tag{B81}$$

While, by singular value decomposition, P can be written as,

$$P = \sum_{i=1}^{N} \sigma_i U_i V_i^T, \tag{B82}$$

where U_i(with size $N \times 1$) and V_i^T (with size $1 \times N$ $\forall i \in \{1, 2, \cdots, N\}$ are singular vectors corresponding to the ith largest singular value. Thus,

$$P \cdot P^T = \sum_{i=1}^{N} \sigma_i^2 \cdot U_i \otimes U_i^T, \tag{B83}$$

where $\otimes$ is the outer product. Therefore, inserting Equation B83 into Equation B81, we have:

$$P' \cdot P'^T \approx \phi^T \cdot P \cdot \phi \cdot \phi^T \cdot P^T \cdot \phi = \phi^T \cdot P \cdot P^T \cdot \phi = \sum_{i=1}^{N} \sigma_i^2 \cdot (\phi^T \cdot U_i) \otimes (U_i^T \cdot \phi) \quad \text{(B84)}$$

This equation holds because $\Phi_i, \forall i \in \{1, 2, \cdots, r\}$ are orthogonal each other, and as a result $\phi \cdot \phi^T = I_{N\times N}$. Therefore, according to Lemma 11, the $1/\alpha$-ordered power of the approximate dynamical reversibility of P' satisfies

$$\Gamma_\alpha^{1/\alpha}(P') \geq ||P'||_F = \mathrm{Tr}(P' \cdot P'^T) \gtrapprox \sum_{i=1}^{r} \sigma_i^2 \cdot \mathrm{Tr}(W_i \otimes W_i^T) = \sum_{i=1}^{r}\sum_{j=1}^{r} \sigma_i^2 \cdot (\phi_j \cdot U_i)^2, \quad \text{(B85)}$$

where, $W_i \equiv \phi^T \cdot U_i = (\phi_1 \cdot U_i, \phi_2 \cdot U_i, \cdots, \phi_r \cdot U_i)^T$, and the second inequality holds because the smallest $N - r$ singular values are cut-off. Thus, if $\phi = (\phi_1^T, \phi_2^T, \cdots, \phi_r^T)$ is selected such that there is at least one vector, say ϕ_j^T, being parallel with the singular vector U_i^T, such that $||W_i||^2$ could be maximized for $\forall i \leq r$. This is possible since both ϕ and U are all mutually orthogonal. This is equivalent to assign more probability mass on the directions of U_i^T for largest singular values.

As a result, $\Gamma_\alpha(P')$ could be maximized. While, according to the approximate relationship $\log \Gamma_\alpha \sim EI$, EI could also be maximized. Therefore, to maximize EI, we should select the coarse-graining strategy that can assign more probability mass on the directions of singular vectors corresponding to the largest singular values.

However, in real applications, the coarse-graining strategy Φ has some constraints (e.g., it should be hard grouping method, and the final TPM of the macro-dynamics should satisfy normalization condition) such that the ϕ_i can not parallel the singular vectors exactly. Thus, this requirement is only an approximate necessary condition for maximizing EI.

B.2 A Numeric Example for Clear Emergence

We will show the consistency between the SVD and the EI maximization methods with two examples. The first example is shown in Figure 1(c) and (d). The TPM is:

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{(B86)}$$

Because $rank(P) = 2$, the clear CE occurs. $EI(P) = 0.81$. We select $r = 2$, and the optimal strategy of coarse-graining for maximization of EI can be written as,

$$\Phi = \begin{pmatrix} 1, 0 \\ 1, 0 \\ 1, 0 \\ 0, 1 \end{pmatrix}, \tag{B87}$$

and the corresponding ϕ being the normalization of Φ is:

$$\phi = \begin{pmatrix} \frac{1}{\sqrt{3}}, 0 \\ \frac{1}{\sqrt{3}}, 0 \\ \frac{1}{\sqrt{3}}, 0 \\ 0, 1 \end{pmatrix}. \tag{B88}$$

The left matrix composed by all the singular vectors of P is:

$$U = \begin{pmatrix} 0 & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{B89}$$

Comparing Equations B88 and B89, we know the first column vectors in Equation B88 are parallel the second and the first column vectors in Equation B89, and the inner products between U_3^T and U_4^T and ϕ are zeros. The final coarse-grained TPM is:

$$P' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{B90}$$

Therefore, $EI(P') = 1$ which is larger than 0.81, and $CE = 1 - 0.81 = 0.19$. This consistent with the conclusion drawn by SVD method where the degree of CE is $\Delta\Gamma = 1/2$.

B.3 An Example for Vague Causal Emergence

We can also give an example of vague CE to illustrate the unreasonable nature of the optimal coarse-graining method for maximizing EI. This example is original shown in [11] to demonstrate the weakness of Erik Hoel's theory of causal emergence.

Consider a Markov chain with 3 states, and its TPM is:

$$P = \begin{pmatrix} 0.3 & 0.6 & 0.1 \\ 0.6 & 0.2 & 0.2 \\ 0 & 0 & 1 \end{pmatrix} \tag{B91}$$

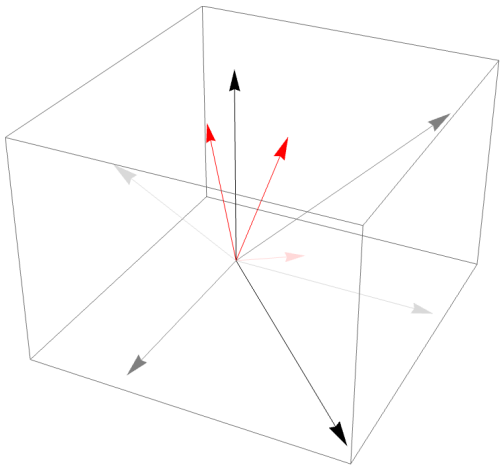

Fig. Supplementary Figure 1 The singular vectors and the vectors representing the coarse-graining strategies. The red arrows are the singular vectors in U(the magnitude is multiplied by the corresponding singular values, the light red arrow represents the singular vector with the smallest singular value), the black arrows represent the optimal coarse-graining strategy with EI maximization, $\{\{1,2\},3\}$. The gray arrows represent the coarse-graining strategy $\{\{2,3\},1\}$, and the light gray arrows represent the strategy $\{\{1,3\},2\}$.

Because $rank(P) = 3$, there is no clear CE. Its singular values are:$\sigma_1 = 1.06, \sigma_2 = 0.80, \sigma_3 = 0.35$. $EI(P) = 0.66$. We can set the vagueness $\epsilon = 0.35$, such that the vague CE occurs with the degree of $\Delta\Gamma = \frac{1.06+0.8}{2} - \frac{1.06+0.8+0.35}{3} = 0.19$.

The U matrix with singular vectors is:

$$U = \begin{pmatrix} 0.32 & -0.67 & -0.67 \\ 0.40 & -0.55 & 0.74 \\ 0.86 & 0.50 & -0.09 \end{pmatrix} \tag{B92}$$

There are totally three possible reasonable coarse-graining strategies: clustering two micro-states as a group and leaving the other micro-state itself as a group, namely, $\{\{1,2\},3\}, \{\{1,3\},2\}, \{\{2,3\},1\}$. The represented vectors for these three strategies, as well as the singular vectors of U can be visualized by Figure Supplementary Figure 1.

It is clearly to see in Figure Supplementary Figure 1 that only the black arrows representing the optimal strategy are parallel with the red arrows which representing the major singular vectors corresponding to the largest two singular values. All other vectors representing the other strategies for coarse-graining all have the non-zero projections on the direction represented by the light red arrow, which is unreasonable.

The optimal strategy is

$$\Phi = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{B93}$$

Finally, the optimal macro-level coarse-grained TPM is:

$$P' = \begin{pmatrix} 0.85 & 0.15 \\ 0 & 1 \end{pmatrix} \tag{B94}$$

It is EI turns out to be 0.68, and the $CE = 0.68 - 0.66 = 0.02$. Thus, the CE occurs according to EI maximization.

However, as pointed out in reference [11], the optimal coarse-graining strategy of EI maximization is unreasonable because it combines the states with dissimilar vectors (the first and the second row vectors), this issue is referred to as ambiguity in [11], as it introduces uncertainty when reducing the intervention from macro-states to micro-states. Also, the coarse-grained TPM P' deviates from P, this can also be observed by the relatively large ϵ.

Another serious problem is the non-commutativity between abstraction(coarse-graining) and marginalization(time evolution) in this example because:

$$P \cdot \Phi = \begin{pmatrix} 0.9 & 0.1 \\ 0.8 & 0.2 \\ 0 & 1 \end{pmatrix}, \tag{B95}$$

but:

$$\Phi \cdot P' = \begin{pmatrix} 0.85 & 0.15 \\ 0.85 & 0.15 \\ 0. & 1. \end{pmatrix}. \tag{B96}$$

Thus,

$$P \cdot \Phi \neq \Phi \cdot P' \tag{B97}$$

Therefore, the coarse-graining operator does not commute with the time evolution operator. That indicates that the optimal coarse-graining method is unreasonable on this example although it can maximize EI.

B.4 Consistency of max EI and SVD Methods in Experiments with Cellular Automata and Complex Networks

This section experimentally demonstrates the relationship between the direction of the clustering method in coarse-graining strategy that maximizes EI, represented by vectors, and the directions of the singular vectors of the TPM, as illustrated in Supplementary Figure 2. The experimental subjects are cellular automata and stochastic block models mentioned in Figures 4 and 3 in the main text for reference. Coarse-graining strategies consist of two steps: 1. identifying the optimal state clustering method using a greedy approach as outlined in [5, 52]; 2. constructing the coarse-grained TPM by summing rows and averaging columns in P based on the clustering method from the first step.

Both figures show that the directions of the vectors representing the states clustering method in coarse-graining strategy that maximizing EI align closely with the singular vectors of the largest retained singular values. Figure Supplementary Figure 2(a) shows the cosine similarities greater than 0.5 between the clustering method vectors and the singular vectors corresponding to the largest singular values. Figure Supplementary Figure 2(b) illustrates the distribution of cosine similarities between the clustering method represented with vectors that maximize EI and the singular vectors, where vec1 (blue column) represents the vectors corresponding to the largest singular values, and vec2 (red columns) represents the vectors corresponding to the

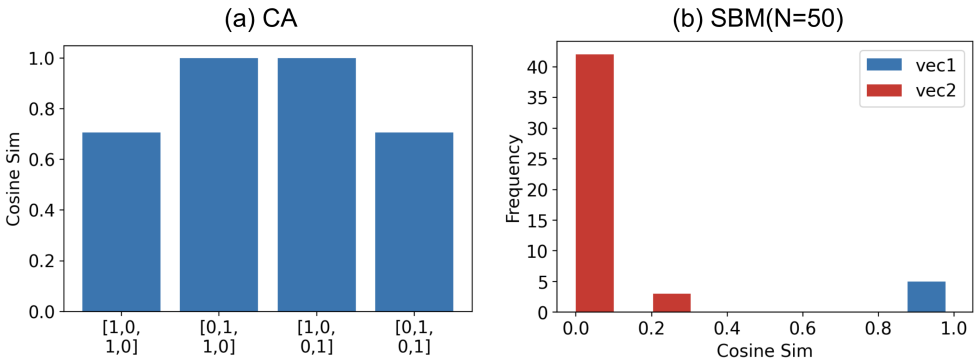

Fig. Supplementary Figure 2 The cosine similarities between state clustering methods in coarse-graining strategies, represented by vectors that maximize EI, and the singular vectors of the TPM for cellular automata (a) and the stochastic block model of complex networks (b) are analyzed. In Figure (a), the four coordinates on the horizontal axis represent the four types of local TPMs (the only possible 2x2 permutation matrices) for the cellular automaton discussed in Figure 4 in the main text. For each local TPM, we identify the coarse-graining strategy—methods for clustering states—represented by vectors that maximize EI using a greedy algorithm ([5, 52]), and then calculate the cosine similarity (the vertical axis) between these vectors and the singular vectors corresponding to the largest r singular values, with r determined by the EI maximization results. Figure (b) illustrates the distribution of cosine similarities between the coarse-graining strategies (methods for clustering nodes into communities) represented by vectors for maximizing EI and the singular vectors for the largest (vec1) and smallest (vec2) singular values for the complex networks randomly generated by stochastic block models (SBM) with 50 nodes, 5 blocks(communities), and parameters $p = 1$ (connection probability within communities) and $q = 0$ (connection probability between communities). We utilize the normalized EI, Eff, as our objective function to compare networks of varying sizes. The horizontal axis represents cosine similarities, while the vertical axis indicates the frequency of these similarities within specific intervals.

smallest values. The similarities for vec1 cluster between 0.8 and 1, while the similarities for vec2 are close to 0. This strongly indicates a high positive correlation between the optimal state clustering directions for EI maximization and the singular vectors of the singular vectors with largest singular values, along with a clear dissimilarity to the ones for smallest singular values.

Supplementary C Experiments on Testing the Correlation between *EI* and Γ

In this section of the Supplementary, we will introduce the details of our experiments on the relationship between the approximate dynamical reversibility Γ and EI on various generated TPMs. The generative model of TPMs have three classes: softening of permutation matrix, softening of controlled degenerative TPMs, and random normalized.

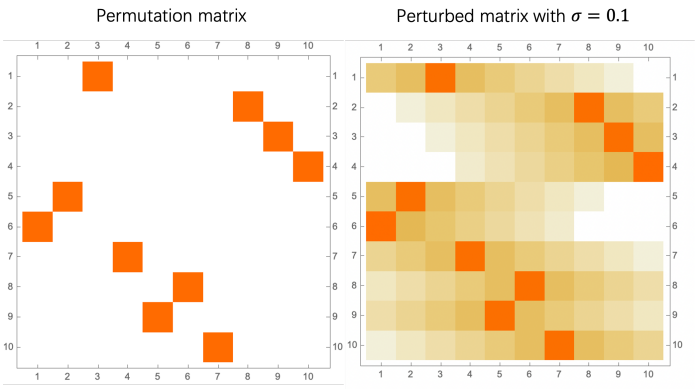

Fig. Supplementary Figure 3 The original TPM which is a permutation matrix and the perturbed matrix after softening.

C.1 Softening of Permutation Matrix

In this series experiments, we will find out what the relationships between Γ and EI are on a variety of TPMs with different deviations from the reversible TPMs (permutation matrix) and different sizes.

For given size N, The TPM is generated by three steps: 1). Randomly sample a permutation matrix with dimension $N \times N$; 2). For each row vector P_i in P, suppose the position of the 1 element is j_i, we fill out all entries of P_i with the probabilities of a Gaussion distribution center at j_i, that is, $p'_{i,j} = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(j-j_i)^2}{\sigma^2}$, where, σ is a free parameter for the degree of softening; 3). Normalize this new row vector P'_i by dividing by $\sum_{j=1}^{N} p'_{ij} = 1$, such that the modified matrix P' is also a TPM.

In this model, the unique parameter σ can control the degree of deviations from the original TPM. When $\sigma = 0$, we recover the original TPM. And when σ increases to very large value, then the row vectors converge to the vector $\mathbb{1}/N$. Supplementary Figure 3 shows the TPMs before and after the update on $\sigma = 10$, where the colors represent the probabilities.

By adjusting different α, we can obtain the similar curves between Γ_α and EI as shown in Supplementary Figure 4. We can observe that the positive correlations between EI and Γ_α can not be changed by different α, the shapes of the curves are different.

C.2 Softening of Controlled Degeneracy

The second model is very similar to the first one, however, the original matrix is not a permutation matrix, but a degenerated matrix. Here, a TPM is degenerative means that there are some row vectors are identical, and the number of identical row vectors is denoted as $N - r$ which is the controlled variable, where r is the rank of P. By tuning $N - r$, we can control the degeneracy [5] of the TPM as Supplementary Figure 5 shows.

Without losing generality, we can start from an identity matrix, and change the controlled $N - r$ row vectors into the same one-hot vectors with all elements 0 except

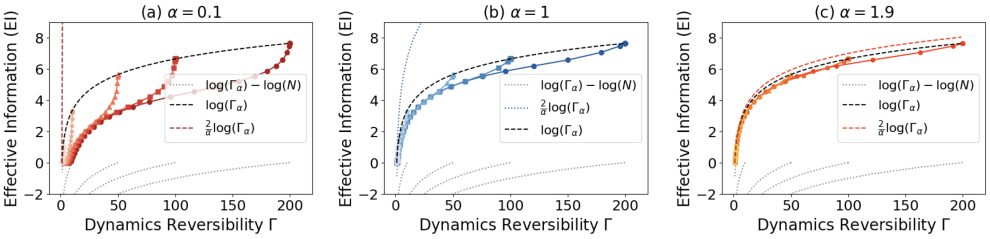

Fig. Supplementary Figure 4 The relationships between EI and Γ_α for different α. The theoretical and empirical upper bounds are shown as red and black dashed lines, while the theoretical lower bounds which change with N are also shown as dotted lines.

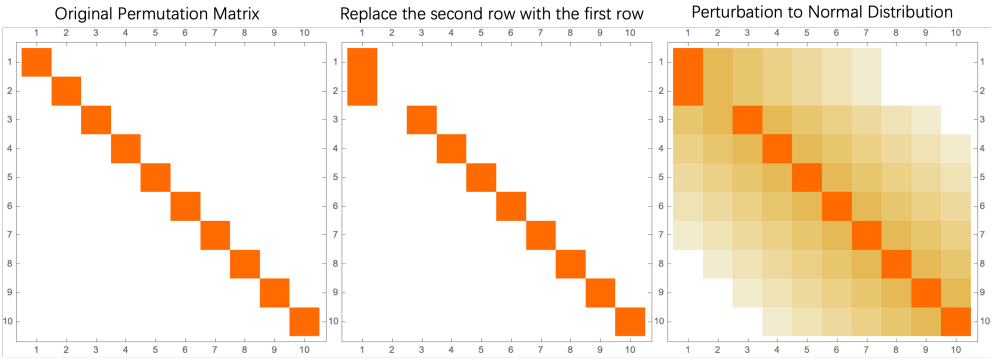

Fig. Supplementary Figure 5 The original TPM ($N = 10$ identity matrix), the replaced degenerated one by controlling $N - r = 2$, and the perturbed TPM of the replaced one by softening with $\sigma = 0.1$. Different colors represent different values of probability.

the first one is 1. After that, we soften the one hot vectors with the same method mentioned in Section C.1 to obtain the results in the main text and Figure 2(b).

C.3 Random Normalization

In this model, only two steps are required to generate a TPM: 1) Sample a row random vector from a uniform distribution in $[0, 1]$, 2) normalize this row vector such that the generated matrix is a TPM.

Supplementary D Analytic Solutions on the Parameterized 2*2 TPM

We compare EI and Γ on a simplest example, the TPM is as:

$$P = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}, \tag{D98}$$

where p and q are all free parameters in the range of $[0, 1]$. With this TPM, we can explicitly write down the expression for EI:

$$\begin{aligned} EI =& \frac{1}{2}\left[p\log_2\frac{2p}{1+p-q}+(1-p)\log_2\frac{2(1-p)}{1-p+q}\right. \\ &\left.+(1-q)\log_2\frac{2(1-q)}{1+p-q}+q\log_2\frac{2q}{1-p+q}\right], \end{aligned} \tag{D99}$$

and Γ:

$$\Gamma = \sqrt{p^2+(1-p)^2+(1-q)^2+q^2+2|1-q-p|}. \tag{D100}$$

According these two expressions, we plot the landscapes in Figure 2(e) and (f).

Supplementary E Applying Our Coarse-graining Method on More Examples

E.1 Examples of Boolean Networks in Hoel(2003)

To compare our methods and EI maximization method on quantification of causal emergence and coarse-graining a Markov chain, we apply our method to the examples of causal emergence in Hoel et al's original papers [5] and [6]. All the examples show clear causal emergence. And almost identical reduced TPMs are obtained for all the examples.

Figures Supplementary Figure 6 and Supplementary Figure 7 show additional examples from the Supplementary, where our coarse-graining method obtain a coarse Markov chain with higher EI compared to Hoel's findings. The main difference lies in: Hoel groups node variables to form a new macro Boolean network and then calculates EI for the TPM. We, on the other hand, group states directly, which might result in a TPM that doesn't fully represent the original Boolean network in terms of variables. Therefore, the extension of our method to variable-based but not state-based coarse-graining method is deserve future studies.

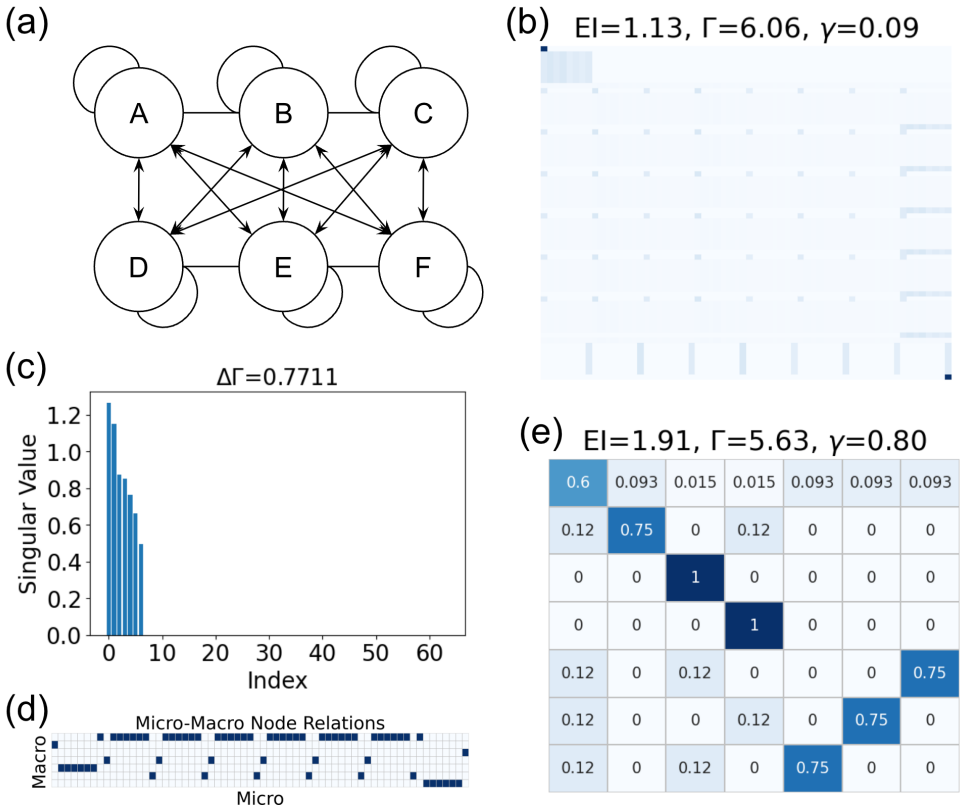

Fig. Supplementary Figure 6 (a) represents a fully connected Boolean network with two distinct types of edges. (b) depicts the Transition Probability Matrix (TPM) as illustrated in Fig. S2 of [5]. (c) displays the singular spectrum of (b). (d) showcases the projection matrix. (e) illustrates the coarse-grained TPM derived from (b) using our coarse-graining approach. In our analysis, we observe a higher macro EI value of 1.91 compared to the 1.84 reported in Hoel et al.'s study. This discrepancy arises from the fact that the macro TPM in [5] encompasses 9 states, whereas our findings involve only 7 states. In their approach, {000111} and {111000} are treated as distinct additional macro states {02} and {20}, whereas we consider them as a single macro-state {11}.

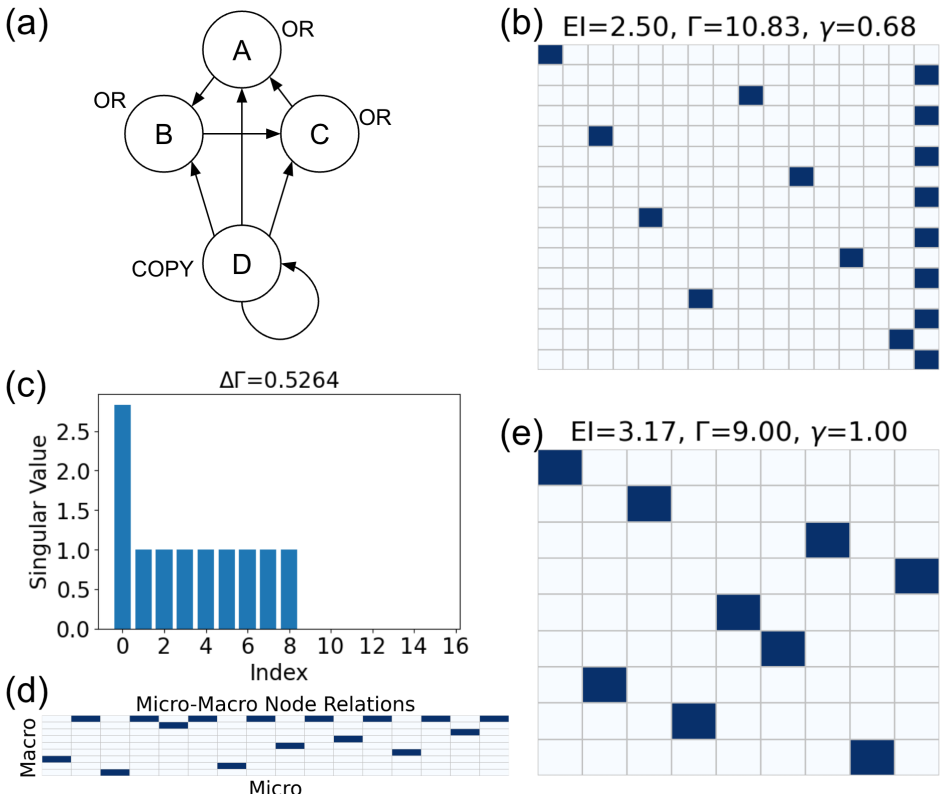

Fig. Supplementary Figure 7 (a) is a directed Boolean network. (b) is the TPM of (a). (c) is the singular spectrum of (b). (d) is projection matrix. (e) is the course-grained TPM. We also get a larger macro EI than Hoel's (3.17 here and 3.00 in [6]). It is because we have 9 macro states and [6] only has 8. We combined all the 8 micro states that transform to {11111111} as a separate macro state, while [6] combines them into each of the other 8 groups.

Supplementary F The Proof of the Commutativity of our Coarse-Graining Strategy

In section 4, we give a coarse-graining method based on the SVD and the stationary distribution. One of the most advantages of this method is that the commutativity of the coarse-graining operator and the time evolution operator(the TPM). In this section, we formalize this characteristic by the following proposition.

Proposition 5. *For a given Markov chain χ with the corresponding TPM P and its stationary distribution μ, and for a given clustering projection matrix Φ defined in 15, we coarse-grain the TPM P according to method mentioned in Section 4 to derive a macro-level TPM represented by P'. Then, the commutativity between the operators of dynamics (P or P') and the coarse-graining (Φ) holds, i.e., the following equation is satisfied:*

$$P \cdot \Phi = \Phi \cdot P'. \tag{F101}$$

Proof. Because μ is the stationary distribution of P, thus,

$$\mu = \mu \cdot P, \tag{F102}$$

And we can define a stationary flow matrix F (Equation 16):

$$F \equiv diag(\mu) \cdot P. \tag{F103}$$

According to Section 4, the coarse-grained stationary flow matrix F' is computed as (Equation 17):

$$F' = \Phi^T \cdot F \cdot \Phi. \tag{F104}$$

The corresponding coarse-grained stationary distribution μ' can be defined as:

$$diag(\mu') = \Phi^T \cdot diag(\mu) \cdot \Phi. \tag{F105}$$

And the reduced TPM P' according to Equation 18 can be written as:

$$P'_i = F'_i / \sum_{j=1}^{r} (F'_i)_j = F'_i / \mu'_i, \forall i \in \{1, 2, \cdots, r\}, \tag{F106}$$

where, the second equality holds because

$$\begin{aligned} \sum_{j=1}^{r} (F'_i)_j &= (F' \cdot \mathbb{1}_{r\times 1})_i \\ &\overset{(F104)}{=} (\Phi^T \cdot F \cdot \Phi \cdot \mathbb{1}_{r\times 1})_i = (\Phi^T \cdot F \cdot \mathbb{1}_{n\times 1})_i \\ &\overset{(F103)}{=} (\Phi^T \cdot diag(\mu) \cdot P \cdot \mathbb{1}_{n\times 1})_i = (\Phi^T \cdot diag(\mu) \cdot \mathbb{1}_{n\times 1})_i = \mu'_i. \end{aligned} \tag{F107}$$

In which, the third equality holds because there is one 1 in each row in Φ and other elements are 0, and the firth equality holds because the sum of each row in P equals 1.

And the matrix form of Equation F106 can be written as:

$$P' = (diag(\mu'))^{-1} \cdot F'. \tag{F108}$$

Therefore,

$$\begin{aligned} P \cdot \Phi &\overset{(F103)}{=} (diag(\mu))^{-1} \cdot F \cdot \Phi \\ &\overset{(F104)}{=} (diag(\mu))^{-1} \cdot (\Phi^T)^{-1} F' \\ &\overset{(F106)}{=} (diag(\mu))^{-1} \cdot (\Phi^T)^{-1} diag(\mu') \cdot P' \\ &\overset{(F105)}{=} (diag(\mu))^{-1} \cdot (\Phi^T)^{-1} \Phi^T \cdot diag(\mu) \cdot \Phi \cdot P' \\ &= \Phi \cdot P' \end{aligned} \tag{F109}$$

□

This shows that the coarse-graining method we have used satisfies:

$$P' = (diag(\mu'))^{-1} \cdot F' = (\Phi^T \cdot diag(\mu) \cdot \Phi)^{-1} \cdot \Phi^T \cdot diag(\mu) \cdot P \cdot \Phi, \tag{F110}$$

and it ensures the commutativity for arbitrary clustering projection Φ.