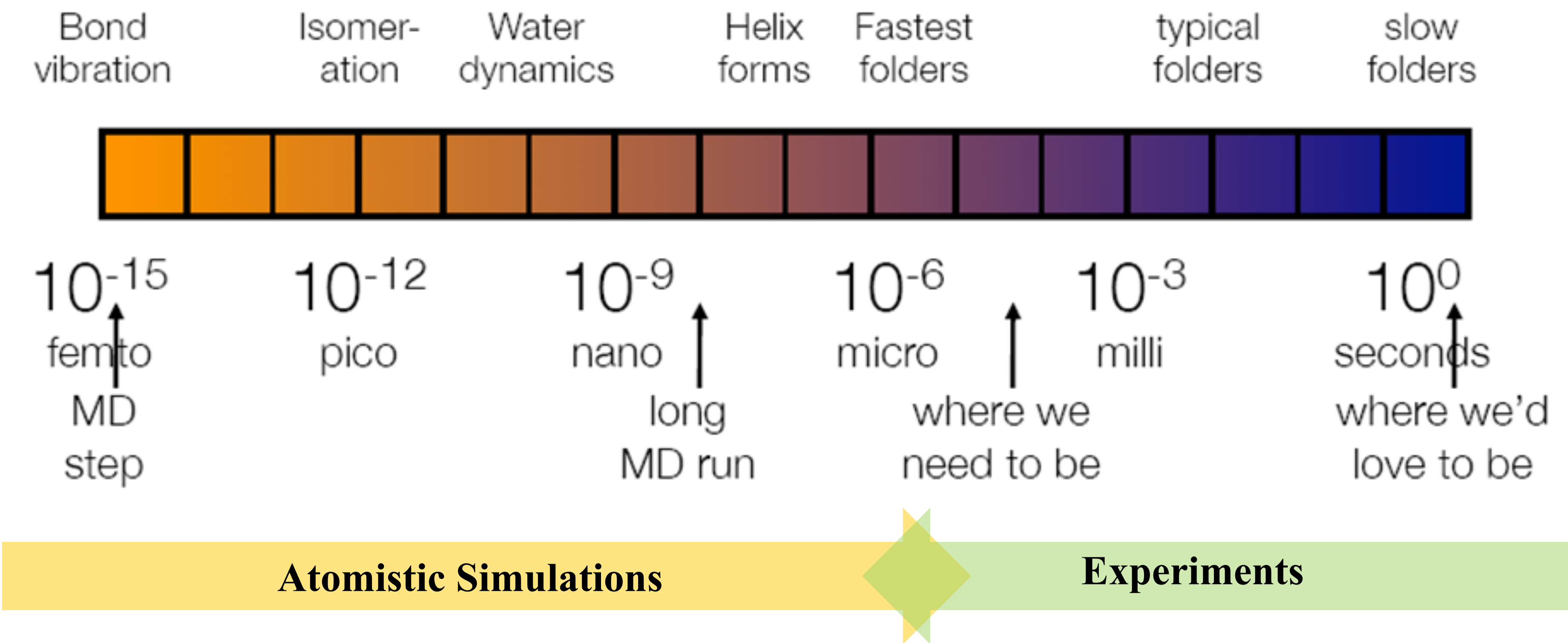
Markov State Models
in
Molecular Dynamics

姚 远

yuany@math.pku.edu.cn

2014.3.29

Key Challenge: Timescale Gap



Solution:

Use short simulations to predict long timescale dynamics

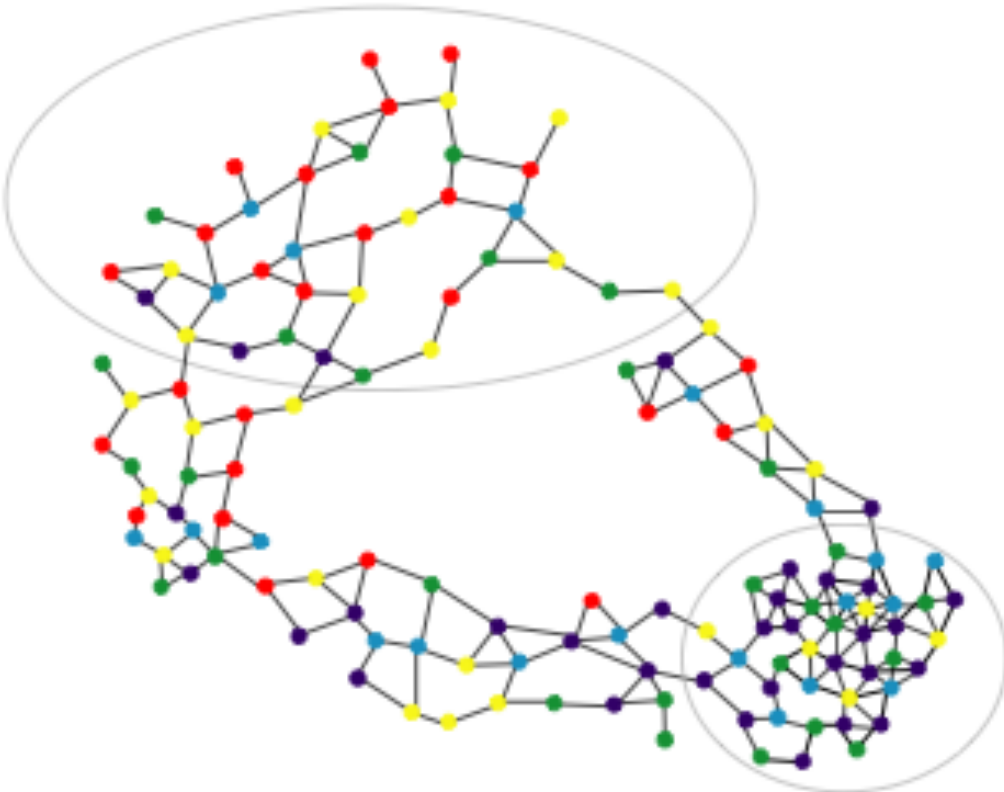
Conformation Network is Too Huge!

Data: A large amount of conformations

Directly work on conformations

Network nodes are snapshots from multiple simulations.

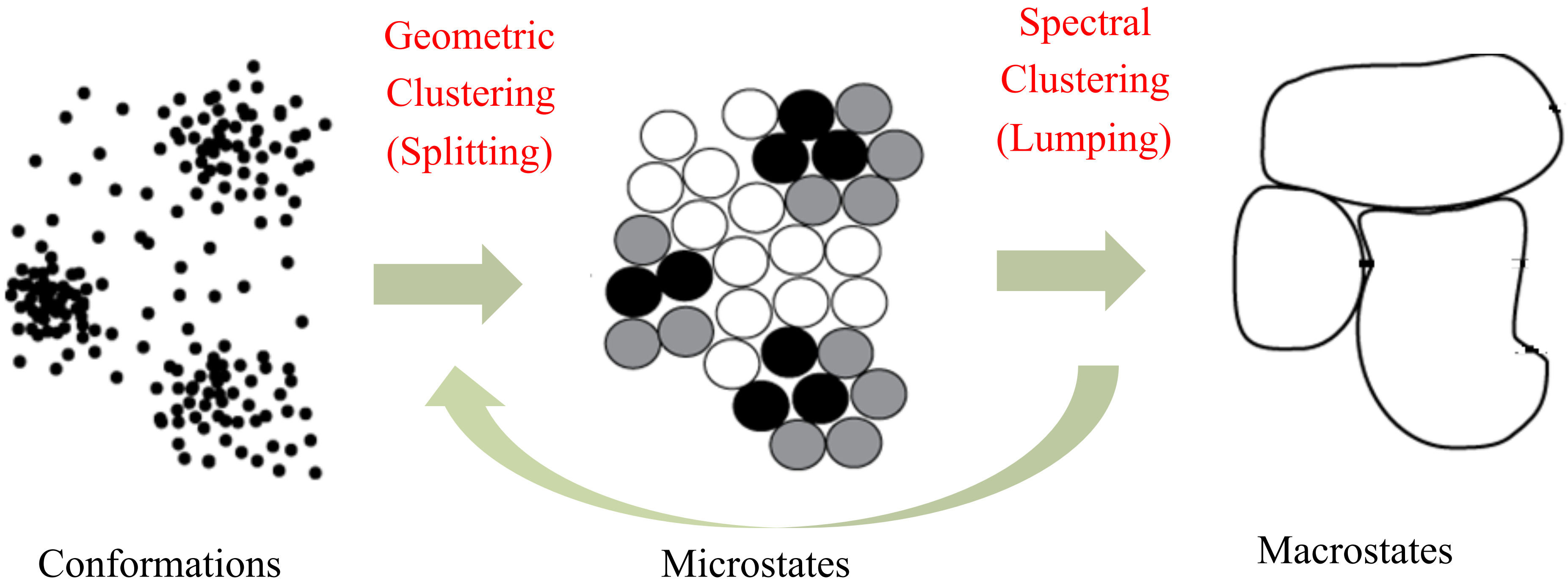
800,000 nodes, 7.4 billion edges



Very Expensive!

Andrec, Felts, Gallicchio & Levy (2005) PNAS, 102, 6801

MSM as Coarse-Grained Approximation

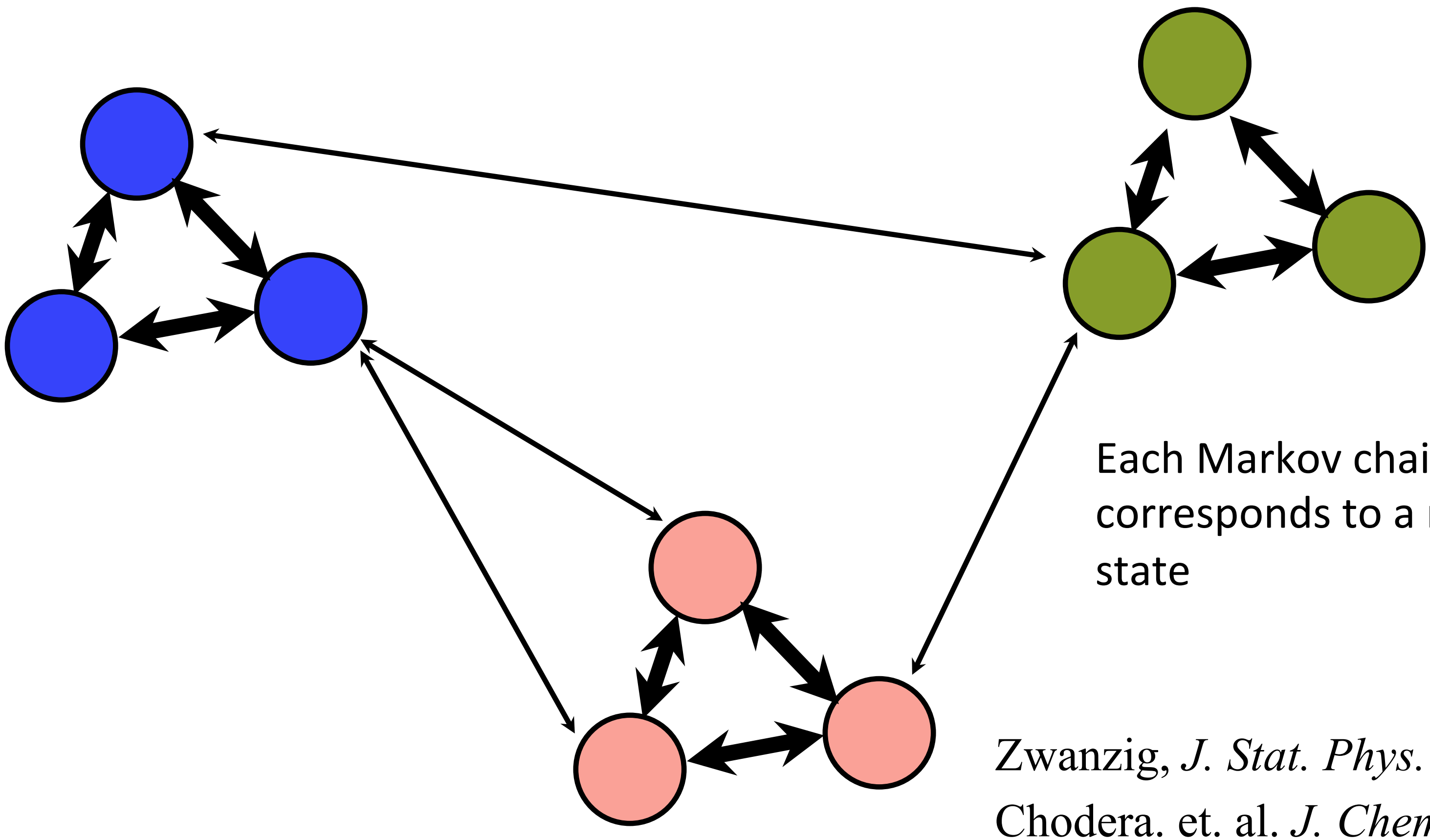


Statistical Inference of MSM

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

Chodera. et. al. *J. Chem. Phys.* 2007
Noé. et.al. *J. Chem. Phys.* 2007
Deuflhard and Weber, *ZIB-report,* 2003
Weber, *ZIB-report,* 2004
Bowman, Huang, and Pande. *Methods* 2009.
Barcalado, et al. *J. Chem. Phys.* 2009

Conformational Dynamics: Nearly Uncoupled Markov Chains



Each Markov chain corresponds to a meta-stable state

Zwanzig, *J. Stat. Phys.* 1983

Chodera. et. al. *J. Chem. Phys.* 2007

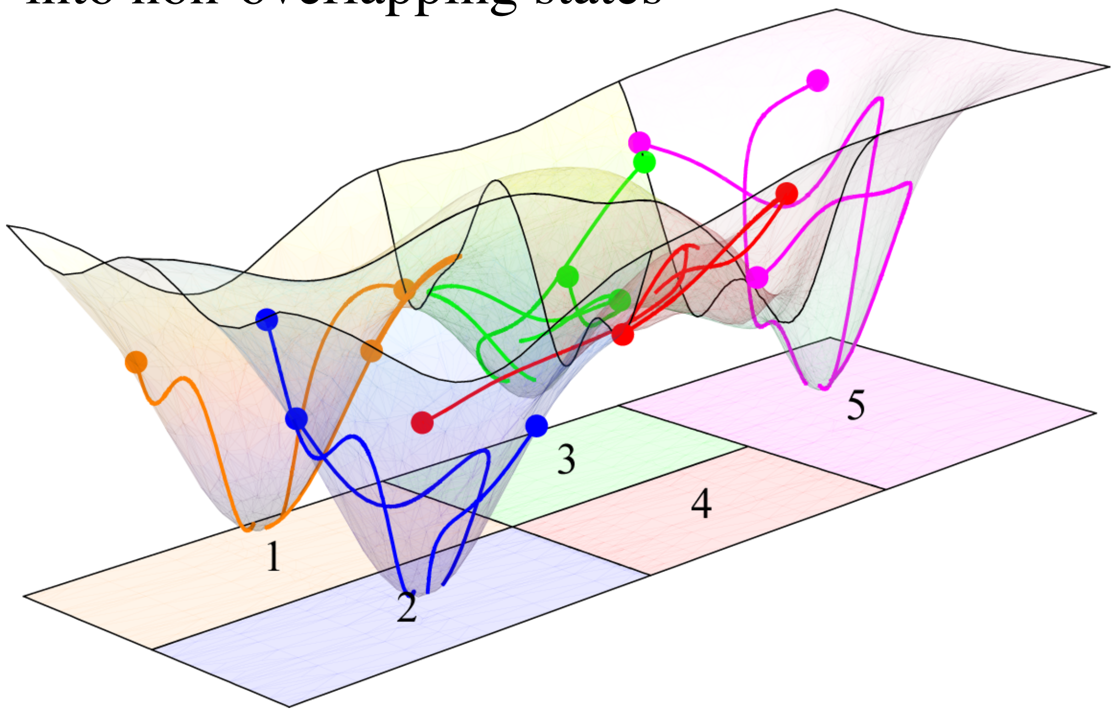
Noé. et.al. *J. Chem. Phys.* 2007

Huang et.al. 2009, Hummer, Shuttle....

Figure Courtesy John Chodera

Free Energy Landscape vs. MSM

The configuration space is decomposed into non-overlapping states



Define transition probabilities between states

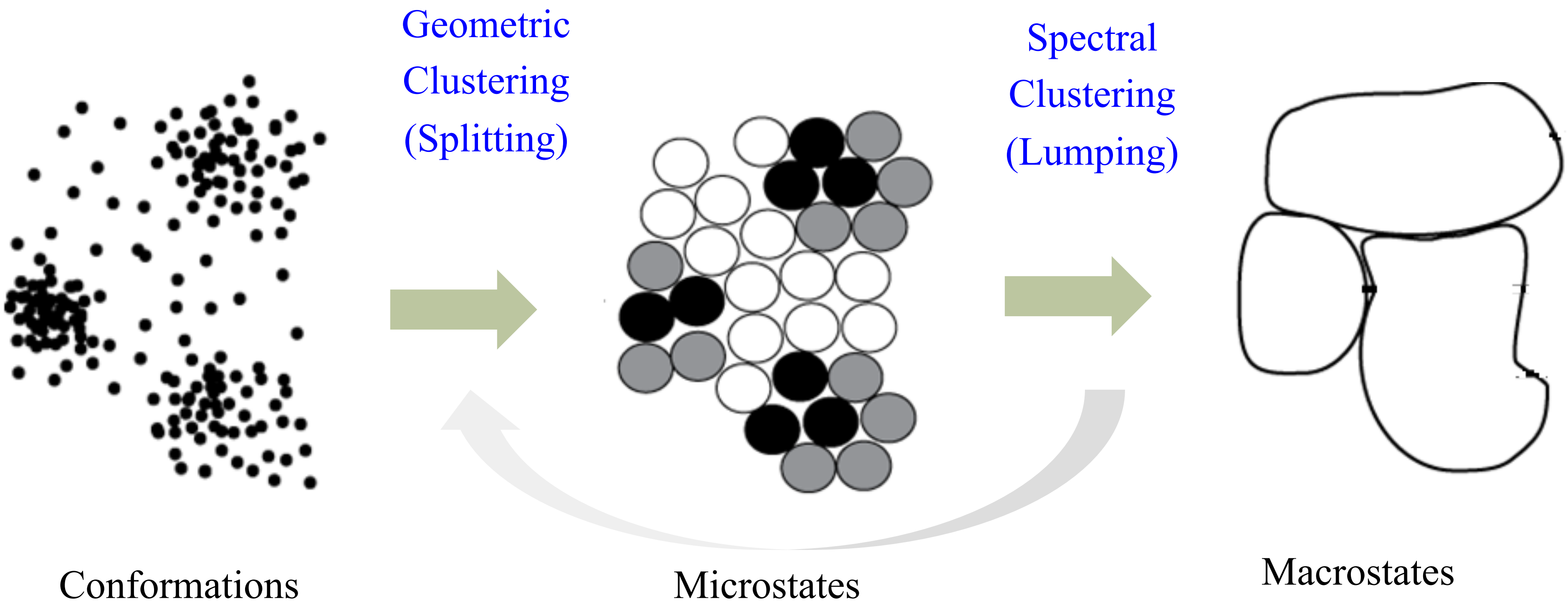
$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

We can extract long time dynamics from MSMs built from short simulations

$$P(n\tau) = [T(\tau)]^n P(0)$$

The time is coarse-grained in τ

Clustering in Biomolecular Dynamics



K-center Clustering with RMSD metric:

Form an epsilon-net to cover the sampled space

Spectral Clustering with Transition Counts:

Find non-spherical metastable states

分子动力系统中的聚类分析

I. Geometric Clustering (距离度量)
- K-means/K-medoids vs. K-center, etc.

II. Kinetic Clustering
- Spectral clustering, etc.

III. 聚类分析的性质
1) Flat clustering vs. Hierarchical clustering
2) Batch vs.Streaming (online) data
3) Complexity and Approximate Algorithms
4) Statistical Consistency

Geometric Clustering based on metric (RMSD)

几种聚类算法比较

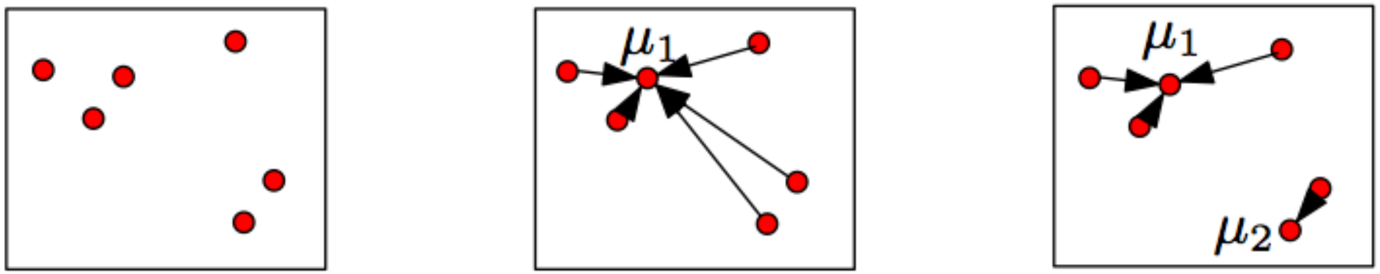
类别	复杂性	近似算法	在线算法	Hierarchical	统计一致性
K-means	NP	50-app	✗	✗	✔[Pollard81]
K-center	NP	2-app. O(kn)	✔ （8-app）	✔ （8-app）	✗ (metric net)
Average-linkage	Close to k-means	?	?	✔	?
Complete-linkage	Close to k-center	a(k)-app k<a(k)<k^log(3)	?	✔	?
Single-linkage	Minimal spanning tree	…	✔ （Persistent Homology）	✔	✔ [Hartigen81,Stuetzle03]

Recall K-center clustering

- input: conformations in a metric space (RMSD) and a number k
- goal: obtain a partition of the points into clusters $C_1, \cdots, C_k$ with centers $\mu_1, \cdots, \mu_k$.
  - condition: minimize the maximum cluster radius:

$$\max_i \max_{x \in C_i} d(x, \mu_i)$$

- NP-hard problem
- 2-approximation algorithm (greedy k-center algorithm)



K-center 几何性质

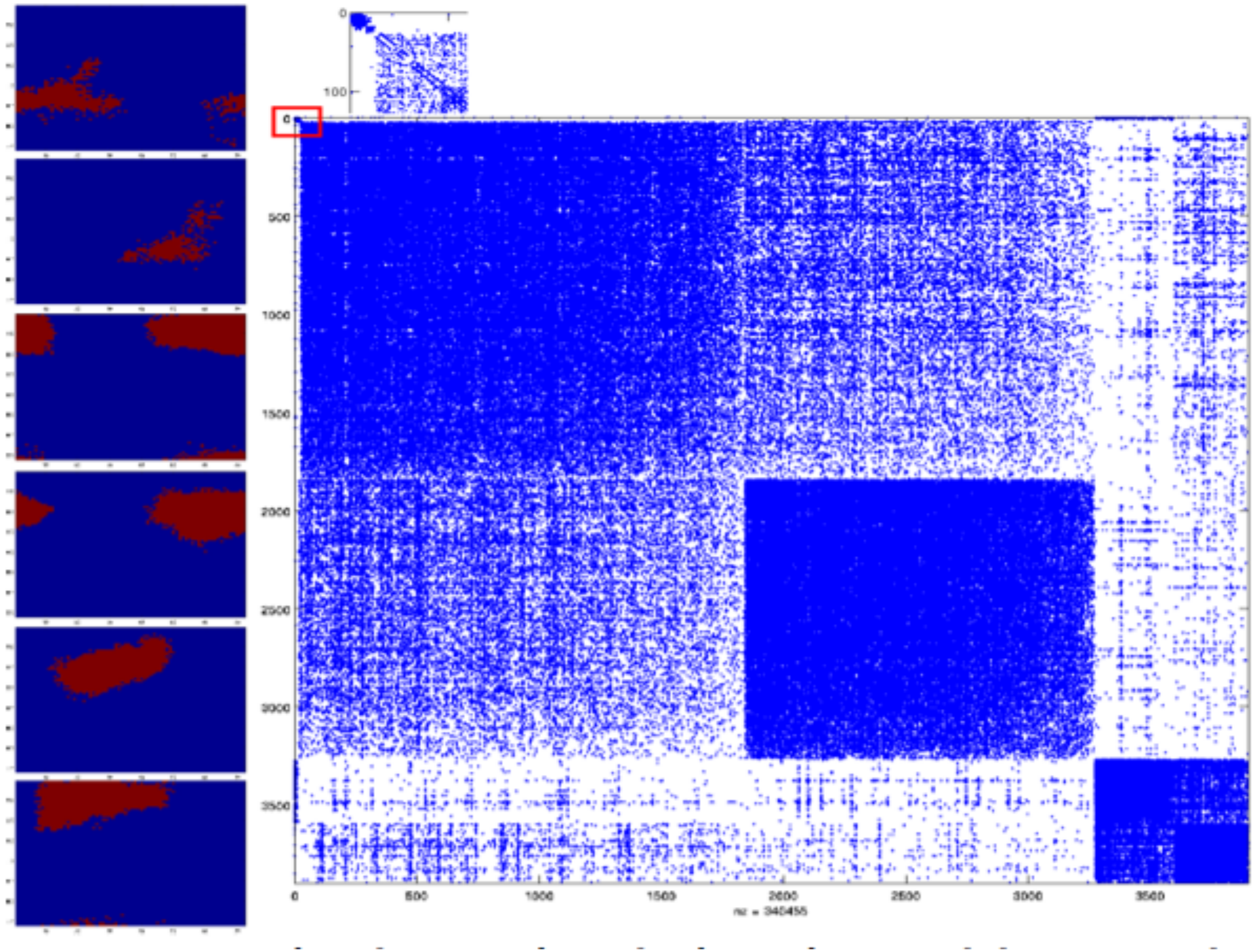
- Farthest-first-traversal算法形成了样本空间的一个度量R-net
  - Any two points in C are R-distance away
  - Points in C form a R-cover of sample space
- K-center is NP-hard, but the 2-approx. algorithm is O(kn), much faster than K-means etc.
- 只依赖于度量结构
- K-center在ISOMAP(TdL'2000, Science)中被采用，称为Landmark技术
- Molecular dynamics application [Sun, Y, Huang, et al. JPC, 09]
- 缺点：
  - 对样本空间边缘的outlier和noise比较敏感 (Good or bad?)
  - 没有statistical consistency theory

Kinetic Clustering

--

Spectral Method

Nearly Block Structure of Transition Matrix



Lumpability of Markov Chains

- Let T be the transition matrix of a Markov chain defined on n states S={1,…,n}.
- P={S_1,…,S_k} is a partition of S into k macrostates.
- Sequences {x_0,…,x_t,…} generated by T, i.e.

$$\mathrm{Prob}(x_t=j \,;\, x_{t-1}=i)= T_{ij}$$

- Induced dynamics: relabel x_t by y_t from corresponding states in partition P
- [Kemeny-Snell'76] T is called *lumpable* if

$$\mathrm{Prob}(y_t=k_0; y_{t-1}=k_1, \ldots, y_{t-m}=k_m) = \mathrm{Prob}(y_t=k_0; y_{t-1}=k_1)$$

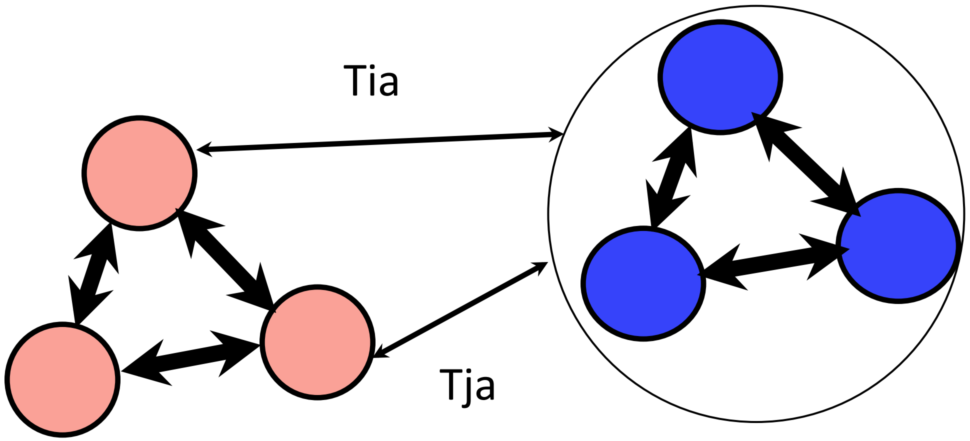
i.e. the induced dynamics is markovian.

Lumpability of Markov Chains

- [Kemeny-Snell'76] T is *lumpable* w.r.t. partition $P=\{S_1,\ldots,S_k\}$ iff for any s, t chosen from P, and for any i, j lying in S_s, the following holds

$$T_{it} = T_{jt}$$

where $T_{it} = \text{sum}_{k \varepsilon St} T_{ik}$.



Spectral Theory of Lumpability

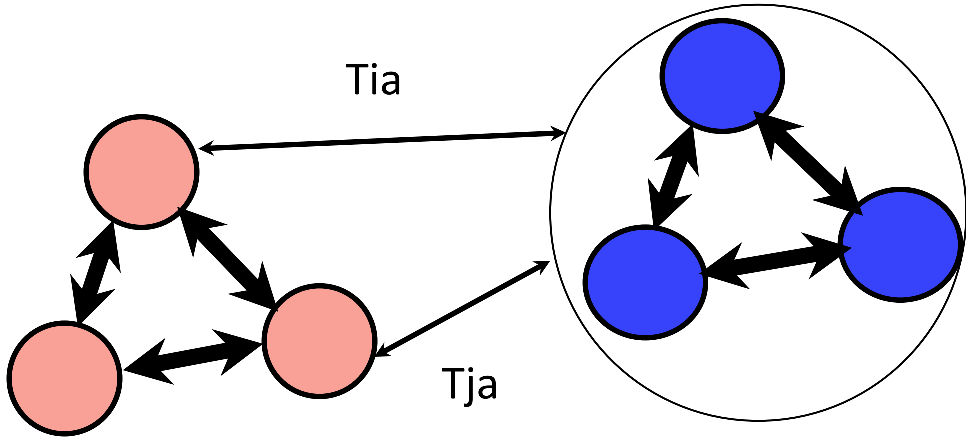
- [Meila-Shi 2001] T is *lumpable w.r.t. P* iff T has k independent piece-wise constant right eigenvectors in the span of characteristic functions of P= $\{S_1,\ldots,S_k\}$.
- Special case: If T is block diagonal, i.e. uncoupled Markov chain, then T is lumpable with piece-wise constant right eigenvectors associated with multiple eigenvalue 1.
- If T is nearly block diagonal, then there are top (k) eigenvectors which fix signs within the block [Belkin-Shi-Yu 2009].
- [E-Li-Vanden_Eijnden 2007] Let T be an n-dim reversible Markov chain, then the best approximation of T from k-dim lumpable chains solves the following optimization

  $$\mathrm{Min}_Q \ \mathrm{norm}(T-Q,\text{`Hilbert-Schmidt'})$$

  where the Hilbert-Schmidt norm of a reversible chain $T = D^{-1}W$, is defined to be sqrt((DT)'(DT))=sqrt(W'W).

Spectral Clustering Theories

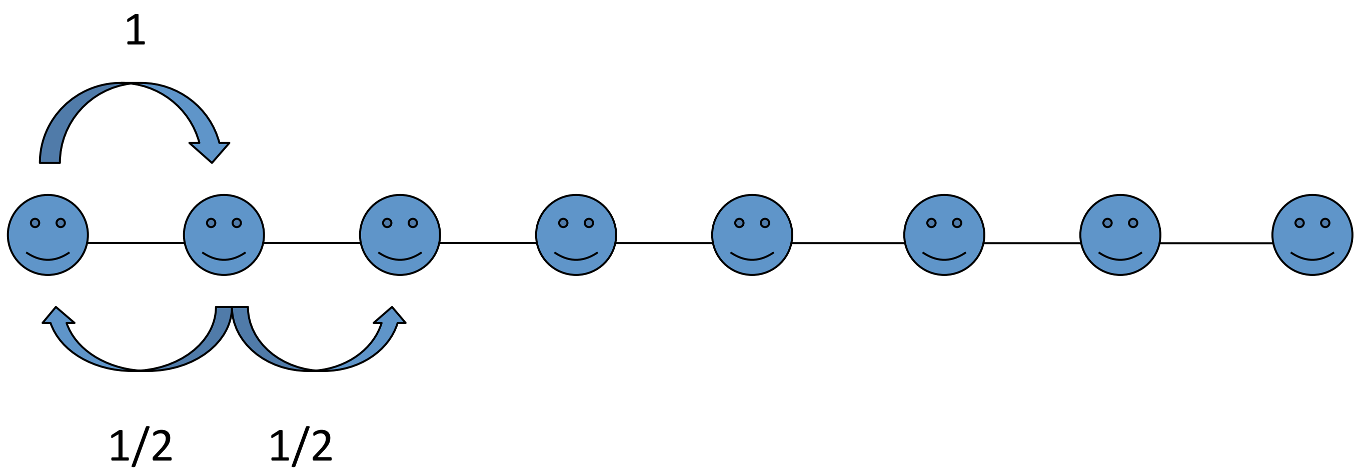
- 3 equivalent descriptions of Lumpability
  - Markovian
  - Spectral properties
    - Piecewise constant r.ev
    - Transition matrix
    - Mean-first-passage



- Approximate Graph min-cut
  - Cheeger's inequalities

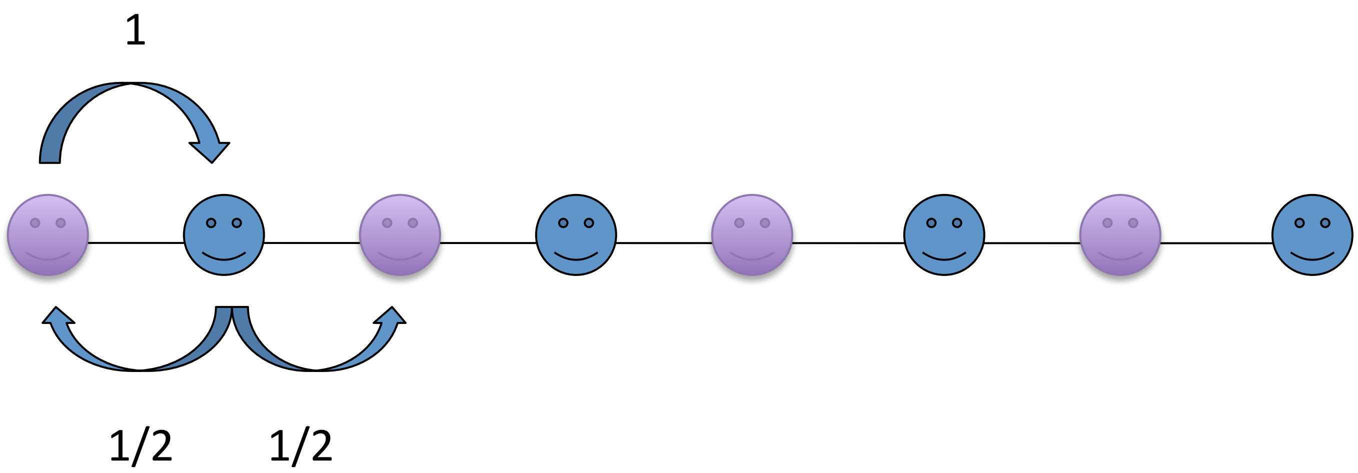
The two theories are different!

Example I



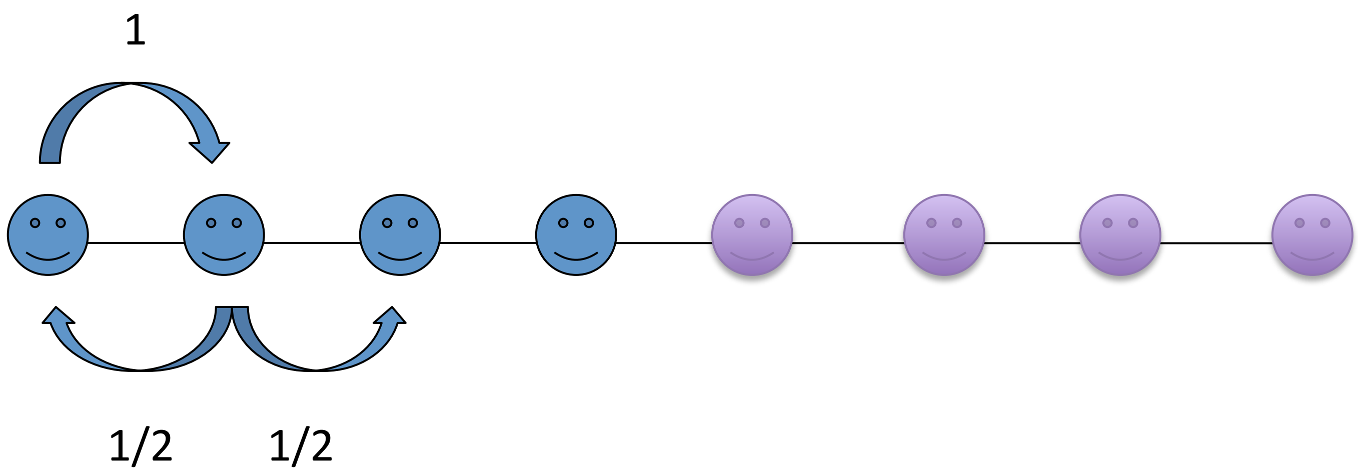
- Consider 2n nodes on a linear chain
- Markov Chain: a node will jump to its neighbors with equal probability
  - $T(i, i-1) = T(i, i+1) = ½$, for $2n>i>1$
  - $T(1,2) = T(2n,2n-1) = 1$

Example I: Lumpable States



- T is lumpable w.r.t. $P^*=(S_{even},S_{odd})$
  - S_{even}: even nodes
  - S_{odd}: odd nodes
- P* corresponds to eigenvector with eigenvalue -1

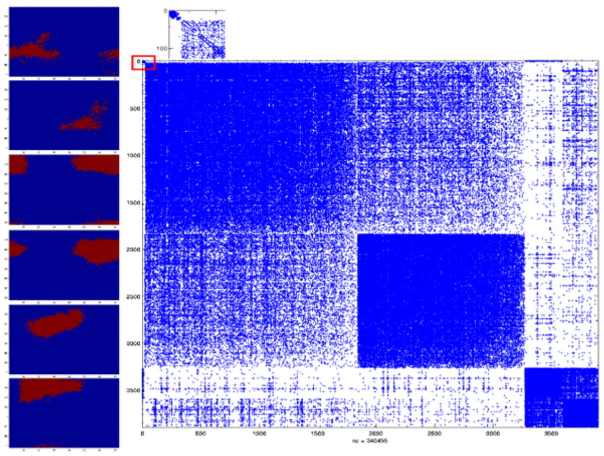
Example I: Graph min-cut



- One graph min-cut given by second largest right eigenvector of T
- n=8,
  - v_2=[0.4714 0.4247 0.2939 0.1049 -0.1049 -0.2939 -0.4247 -0.4714]
  - Eigenvalue is 0.9010

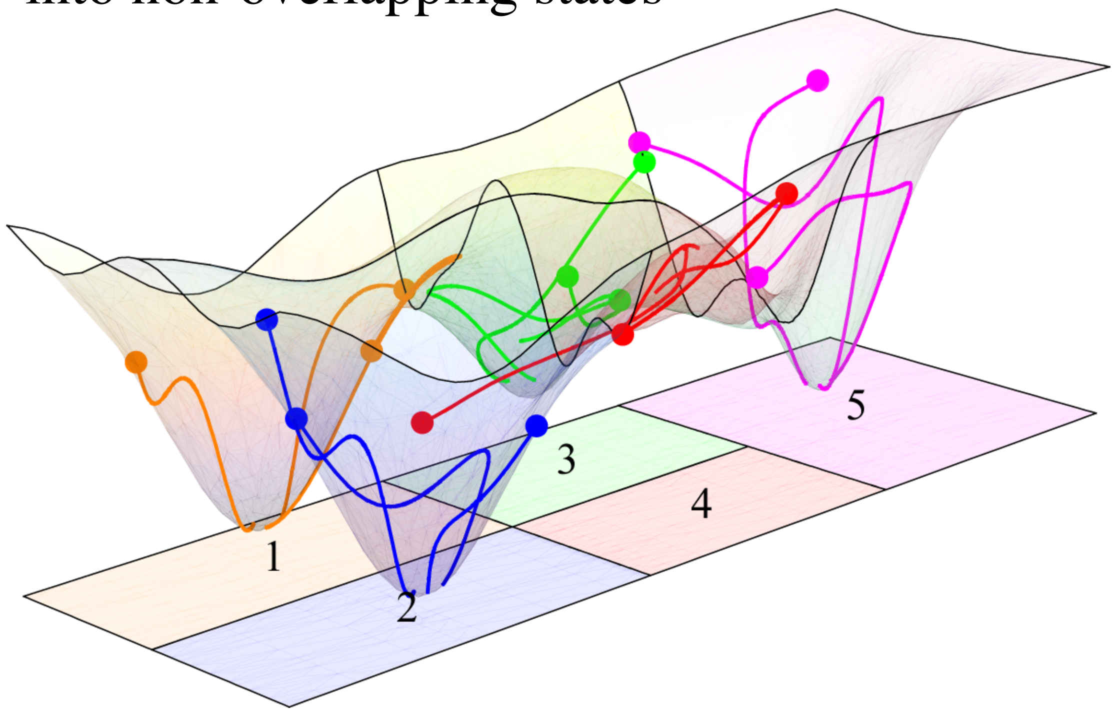
When two theories meet?

- [Meila-Shi 2001, E-L-V 2008]
  - If the top k eigenvectors are piecewise constant functions w.r.t. partition $P=\{S_1,\ldots,S_k\}$
  - Or, T is nearly uncoupled Markov chain (nearly block diagonal)



Free Energy Landscape drives MSM

The configuration space is decomposed into non-overlapping states



Define transition probabilities between states

$$T(\tau) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{15} \\ p_{21} & p_{22} & & \\ \vdots & & \ddots & \\ p_{51} & & & p_{55} \end{bmatrix}$$

We can extract long time dynamics from MSMs built from short simulations

$$P(n\tau) = [T(\tau)]^n P(0)$$

The time is coarse-grained in τ

Spectral Clustering Algorithm

- Typical spectral algorithm to find approximate lumpable states in nearly uncoupled systems [Ng-Jordan-Weiss'02]:
  - Find top k right eigenvectors of T where a large spectral gap occurs, $v_1,\dots,v_k$
  - Embed the data into R^k by those eigenvectors
  - Use k-means (or alternatives) to find k clusters in R^k
- In biomolecular dynamics, this type algorithm is named after Perron, or PCCA [Weber'04].

Problems

- Standard spectral clustering algorithms may fail due to
    - Sparsely sampled microstates are isolated
    - Discovered as spurious metastable states
- Solutions:
    - Hierarchical/Multiresolution Nystrom method ([Huang et al. 2010, Yao et al. 2013])
    - Other non-spectral methods? Yes, Milestoning ([Schutte et al. 2011])

Statistical Inference of MSM

- Maximum Likelihood
- Bayesian Inference of Reversible Markov Chains [Bacallado et al. 2011, 2013]

Analysis of MSM

- What can we do with a discrete Markov State Model?
  - Mean-first-passage-time from state a to state b
  - Transition path theory: reaction current (flux) from source set A to sink set B
    - Continuous space [E and Vanden-Eijden, 2006]
    - Discrete space [Metzner et al. 2009; Noe et al. 2009]
  - Topological landscape [E, Lu and Yao, 2014]

Reference

- Shi, Belkin, and Yu, Data spectroscopy: Eigenspaces of convolution operators and clustering. Annals of Statistics, 37 (6B): 3960-3984. 2008.
- Chodera, J. D., Singhal, N., Pande V. S., Dill, K. A., and Swope W. C. (2007) *J. Chem. Phys.,* ***126,*** 155101**-.**
- E, Li, and Vanden_Eijnden. Optimal partition and effective dynamics of complex networks. PNAS, 105 (23): 7907–7912, 2008.
- T.F. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:293-306, 1985.
- J.A. Hartigan. Consistency of single linkage for high-density clusters. Journal of the American Statistical Association, 76:388-394, 1981.
- Kemeny and Snell 1976. Finite Markov Chains. Springer-Verlag.
- Meila and Shi, A random walk view of spectral segmentation, AISTATS 2001.
- Andrew Y . Ng, Michael I. Jordan and Yair Weiss. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems (NIPS) 14, 2002.
- D. Pollard. Strong consistency of k-means clustering. Annals of Statistics, 9(1):135-140, 1981
- W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. Journal of Classification, 20(5): 25-47, 2003.
- Deuflhard, P. and M. Weber, Robust Perron Cluster Analysis in Conformation Dynamics, ZIB-Report 03-19, 2003.
- Weber, M. Clustering by using a simplex structure, ZIB-Report 04-03, 2004
- Ulrike von Luxburg, A Tutorial on Spectral Clustering, Max Planck Institute for Biological Cybernetics, TR-149, 2006
- Huang, X., Y. Yao, J. Sun, L. Guibas, G. Carlsson and V.S. Pande. Constructing Multi-Resolution Markov State Models (MSMS) to Elucidate RNA Hairpin Folding Mechanisms. Proceedings of the Pacific Symposium on Biocomputing, 15, 228-239, 2010.
- Yuan Yao, Raymond Z. Cui, Gregory R. Bowman, Daniel Silva, Jian Sun, Xuhui Huang. Hierarchical Nystrom Methods for Constructing Markov State Models for Conformational Dynamics. *J. Chem. Phys.,* 138 (17):174106. arXiv:1301.0974, 2013.
- C Schütte, F Noé, J Lu, M Sarich, E Vanden-Eijnden (2011). Markov state models based on milestoning. *J Chem Phys*., 134 (20): 204105
- Weinan E and Eric Vanden-Eijnden, Towards a theory of transition paths, *J. Stat. Phys.* 123 (2006), 503–523.
- Weinan E and Eric Vanden-Eijnden, Transition-path theory and path-finding algorithms for the study of rare events, Annual Review of Physical Chemistry 61 (2010), 391–420.
- Philipp Metzner, Christof Schu¨tte, and Eric Vanden-Eijnden, Transition path theory for markov jump processes, Multiscale Model. Simul. 7 (2009), 1192.
- F. Noe, C. Schu¨tte, E. Vanden−Eijnden, L. Reich, and T. R. Weikl, Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations, Proceedings of the National Academy of Sciences of the United States of America, 106:45 (2009), pp. 19011–19016.
- Sergio A. Bacallado. Bayesian analysis of variable-order, reversible Markov chains. The Annals of Statistics, (39), 2, pp. 838-864, 2011.
- Sergio A. Bacallado, Stefano Favaro, Lorenzo Trippa. Bayesian nonparametric analysis of reversible Markov chains. The Annals of Statistics, (41), 2, pp. 870-896, 2013.

Some material at -- http://www.math.pku.edu.cn/teachers/yaoy/Spring2011/