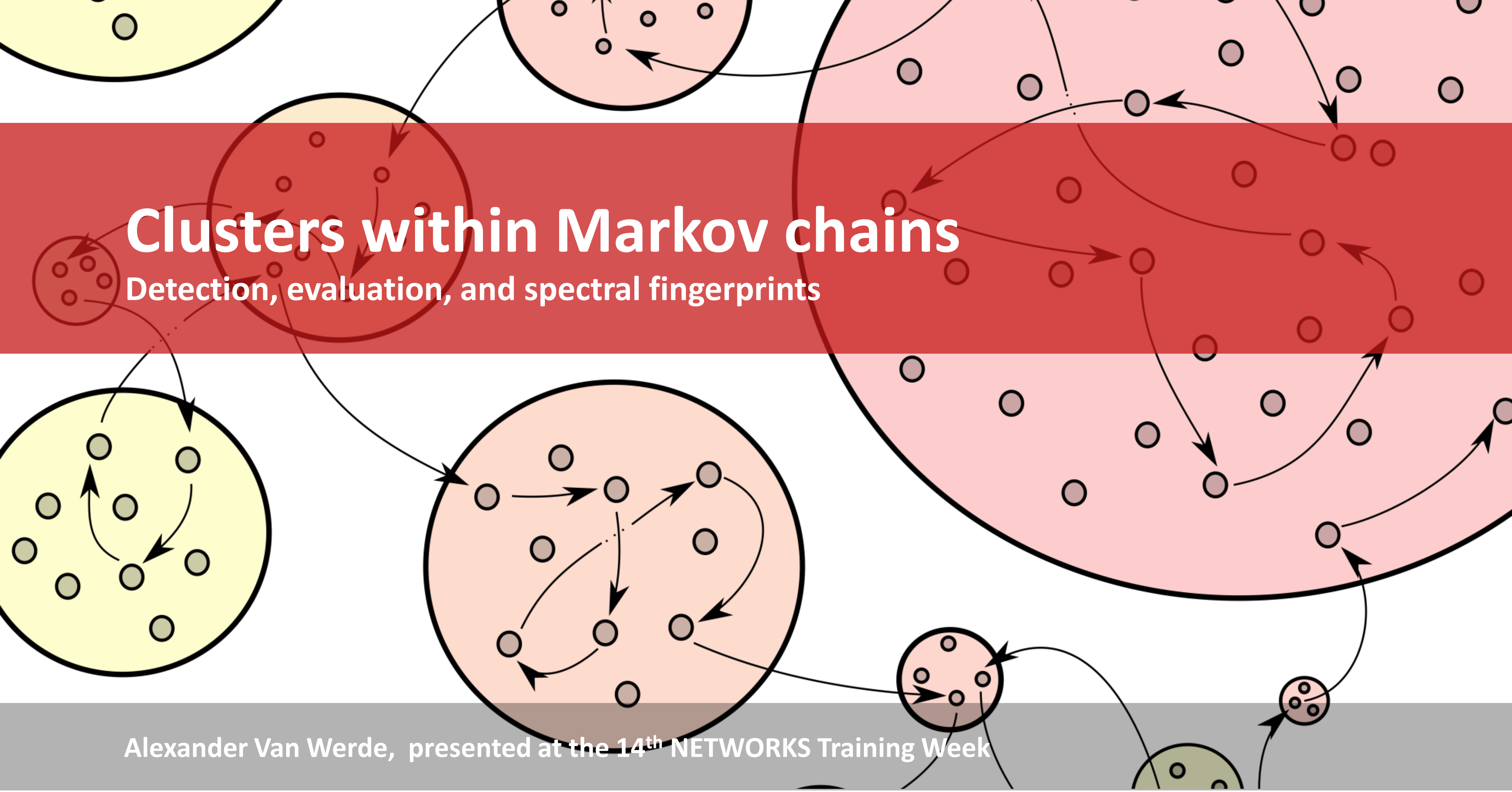


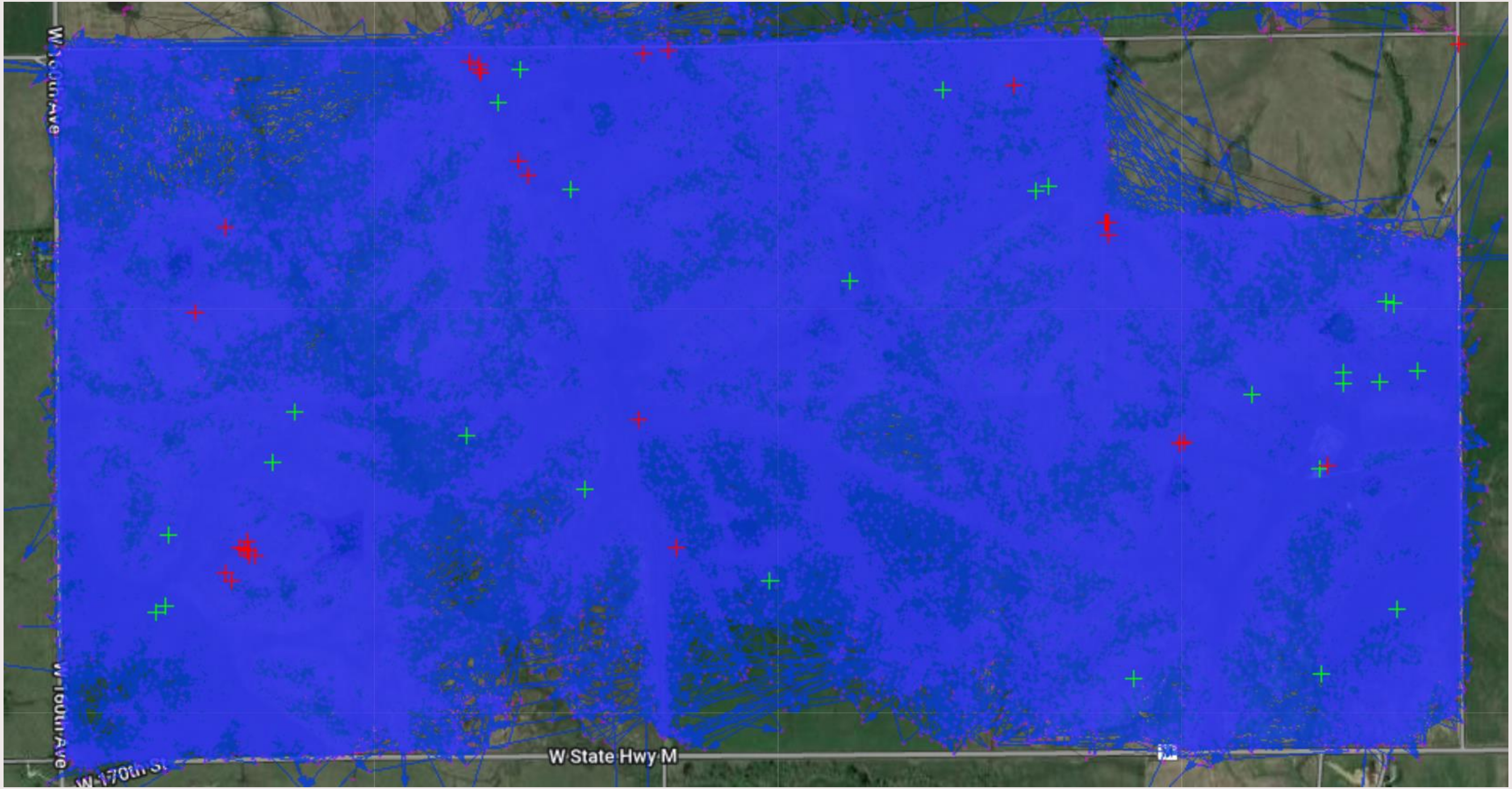
Clusters within Markov chains

Detection, evaluation, and spectral fingerprints

Alexander Van Werde, presented at the 14th NETWORKS Training Week

Joint work with Albert Senen-Cerda, Gianluca Kosmella, and Jaron Sanders

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

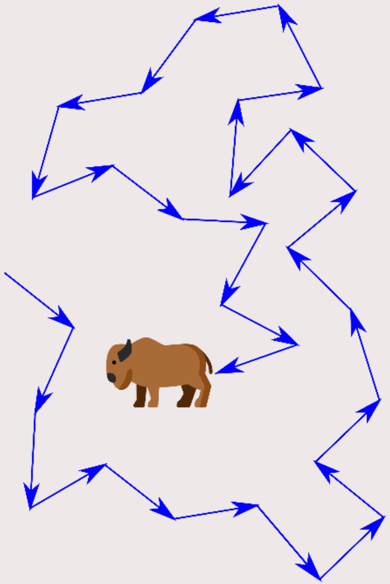


Dunn ranch bison movements. Collected by Stephen Blake, Randy Arndt, and Doug Ladd with the Max Planck Institute in collaboration with the Natural Conservancy (Missouri).

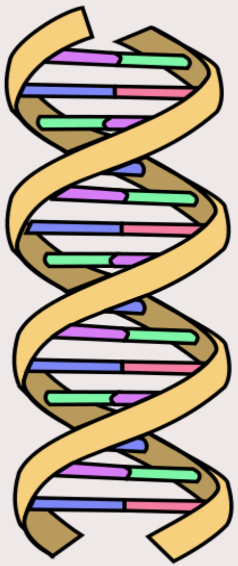
Goal: To understand role of bisons in ecological restoration of prairies and to *understand bison herd dynamics*.

Many processes can be modelled as Markov chains

Animal movement



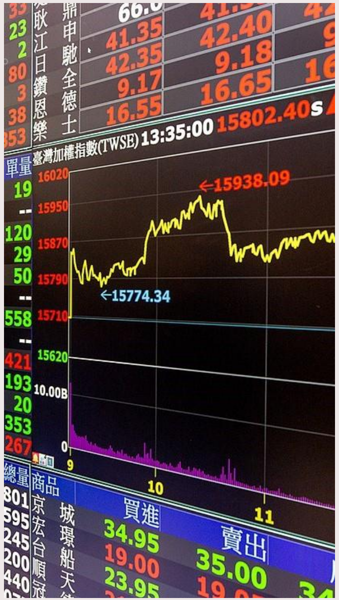
DNA



Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas in turpis sit amet risus ultricies tincidunt vitae accumsan enim. Pellentesque ac efficitur enim. Praesent dui magna, mattis at bibendum eu, fringilla eu neque. Nulla vel accumsan arcu. Quisque sagittis, massa quis maximus vestibulum, sapien ligula venenatis ex, eu elementum diam ante a neque. Nunc egestas malesuada interdum. Morbi eget sem in neque pretium vehicula. Duis nulla turpis, efficitur sed lacus at, auctor tempor lorem. Maecenas pharetra et erat non viverra. Morbi pellentesque velit nec risus imperdiet, ac fermentum lectus aliquam. Maecenas rutrum nisi ut turpis ultricies, id placerat quam fringilla. Donec non interdum urna. Nam justo mi, malesuada eu sollicitudin vitae, euismod vitae tortor. Donec feugiat ligula sed sem luctus consectetur id non lectus. Donec ac neque sed metus facilisis rhoncus. Pellentesque iaculis, urna et consequat venenatis, nulla felis laoreet metus, vitae porta risus neque non odio. Sed venenatis mauris magna, sed tincidunt enim suscipit vitae. Donec a commodo ipsum. Proin euismod lacus ac metus finibus sagittis. Duis congue lorem quis velit tempus tincidunt. Ut ultrices rhoncus ipsum, eget aliquam ex dapibus et. Mauris porta, quam sit amet tincidunt aliquet, ex ante tincidunt nibh, et ultricies mi diam ac odio. Duis viverra velit ut dolor ornare ultricies. Phasellus non interdum velit, ac egestas ligula. Cras magna odio, vestibulum in ex ut, scelerisque dapibus tellus. Curabitur facilisis nisl quis erat lobortis efficitur. Donec at ullamcorper nulla, ut eleifend purus. Donec euismod odio quis dui ultricies efficitur. Quisque lobortis lectus tortor, in interdum lorem tempor dictum. Curabitur vitae libero sed ex elementum fringilla. Duis vel sagittis nisl. Donec a lectus consectetur quam tincidunt mollis venenatis eget purus. Mauris condimentum id dolor quis rutrum. Etiam egestas sapien sit amet interdum tincidunt. Nam a viverra risus. Nulla in auctor massa, ac lobortis nunc. Nam maximus efficitur purus a laoreet. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos …

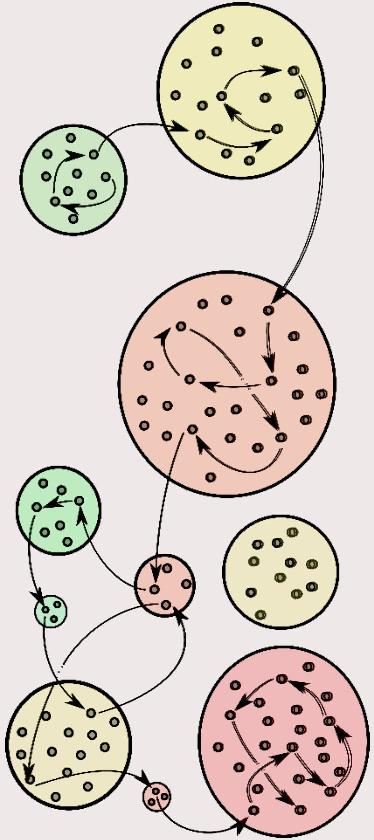
Stock market



TU/e

Problem

- State space of real-world Markovian models is large
  - ➢ Human interpretation is difficult
  - ➢ Algorithms slow down
  - ➢ Overfitting
- Can we identify a lower-dimensional structure?
  - ➢ Clustering
- Does the theory apply to the real world?
- How to evaluate?



Some related work

Community detection in graphs: Stochastic block model

Holland, P.W., Laskey, K.B. and Leinhardt, S., *"Stochastic blockmodels: First steps."* Social networks, 1983.

E. Abbe, *"Community Detection and Stochastic Block Models: Recent Developments,"* Journal of Machine Learning Research, 2017.

Random matrices with dependence

Banna, Marwa, Florence Merlevède, and Magda Peligrad. *"On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries."* Stochastic Processes and their Applications, 2015

Clustering in reinforcement learning

O.-A. Maillard and S. Mannor, *"Latent bandits.,"* in International Conference on Machine Learning, 2014.

Jedra, Y., Lee, J., Proutière, A., and Yun, S. Y.*"Nearly Optimal Latent State Decoding in Block MDPs."* arXiv preprint arXiv:2208.08480, 2022.

Hidden Markov Models

R. S. Mamon and R. J. Elliott, *"Hidden Markov Models in Finance."* 2007.

C. Manning and H. Schutze, *"Foundations of Statistical Natural Language Processing."* 1999

Model evaluation of Markov chains

P. Billingsley, *"Statistical methods in Markov chains,"* The Annals of Mathematical Statistics, 1961.

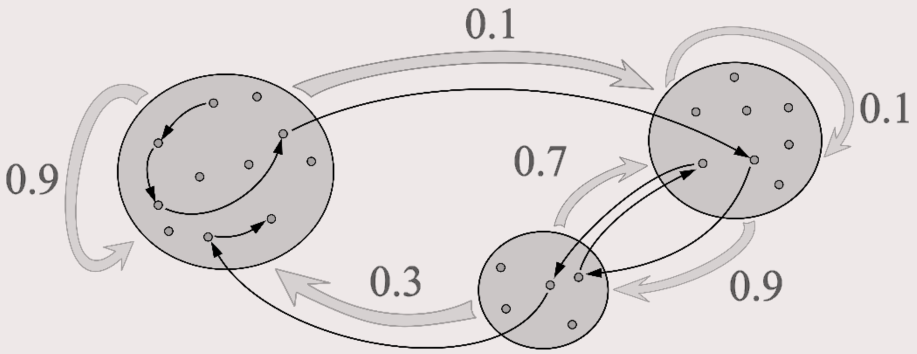
See Section 4 in our paper *"Detection and Evaluation of Clusters within Sequential Data."* arXiv:2210.01679, (2022) for a more in-depth review of the literature.

Block Markov chains

Definition. Fix $K \geq 1$ and a stochastic matrix $p \in [0,1]^{K \times K}$. Let $\mathcal{V}_1, \dots, \mathcal{V}_K$ be a partition of $\mathcal{V} := \{1, \dots, n\}$.

The associated *block Markov chain* is the Markov chain on $\mathcal{V}$ with transition matrix

$$P_{i,j} = \frac{p_{x,y}}{\#\mathcal{V}_y} \qquad \forall i \in \mathcal{V}_x,\ \forall j \in \mathcal{V}_y.$$



Question. Given a sample path $X_0, \dots, X_\ell$. Can you recover the clusters $\mathcal{V}_1, \dots, \mathcal{V}_K$?

Answer. Yes!

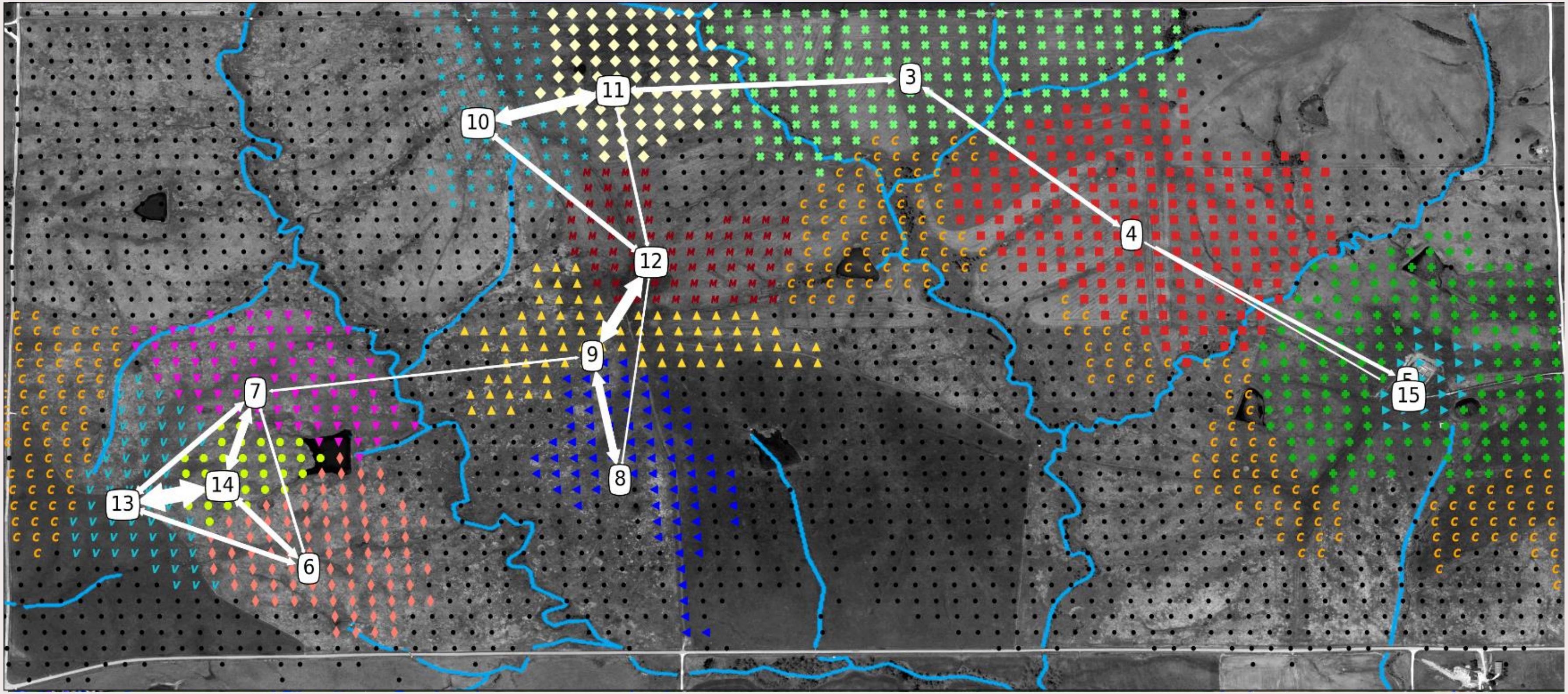
Algorithm. (Sanders et. al., 2020)

1. Build a matrix $\widehat{\boldsymbol{N}}_{i,j} := \#\{\text{transitions i} \to \text{j}\}$
2. Remove noise: Singular value decomposition yields low-rank approximation.
3. Initial guess: K-means algorithm
4. Improvement algorithm: local maximization of the log-likelihood function.

Question. Does the algorithm work in the real world?

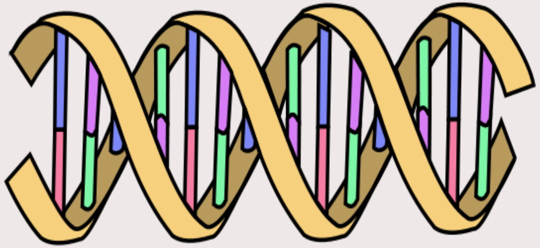
Question. Real-world data does not come with a ground truth. How do we evaluate?

Clusters in animal movements



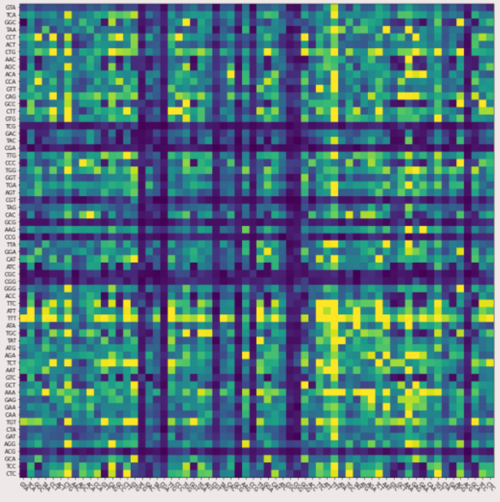
Clusters in DNA

```
ATA CTA ATT TTA TAC AGT TTT CAG GGA TTT
TTT TTC CTC CCT GAG CAA CAA ATT GAT GTA
TAT ATT TGA AGG AGC TAA TAC AAC TTT TTG
CAA CCA ATA TGA AAG TAC CAA TTT TTA AAA
ATA ATT GCT AAA ATT CTT TAT ATA TTT TGT
TAA GGA ATT AAT AAT ATA CCA TTC AAC TTG
AAA AAT GAT GTA TAC AGA AGA AAA TAT GTA
CTT TGG GAA TGT GTA ACT GTT GGG ATT TTT
CTC TTT TCC TCA TTG TCC TTT AAG TAT GAC
GCA AAT TAT TTT TTA GAG GTA AAA TAT GGC
ATG CTT TGG AGA TAT TGC AGA TTC AGC TTC
AAA CCA CCA CAA TAA AGC AAA TAT TAC AAT
AGT GAG TCA CAC AAA ATT TTT TTT TGT TTC
CCA GTT CAT ATA AAA TTT ATA CTT ACA CTA
TAC TGT AGT CTA AGT ...
```

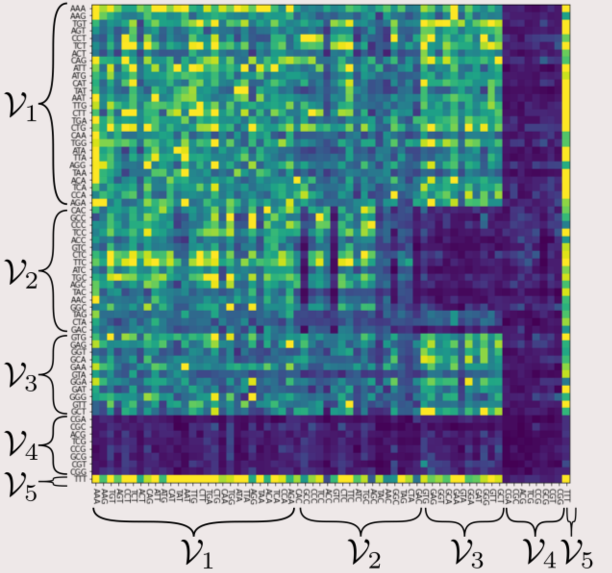


Recall: $\widehat{N}_{i,j} \coloneqq \#\{\text{transitions i} \to \text{j}\}$

Before clustering:



After clustering:



Clusters in text

200 clusters including:

$\mathcal{V}_{\text{color}}$ = {green, white, red, blue, black, gold}

$\mathcal{V}_{\text{letters}}$ = {b, c, x, v, iii, g, e, r, f, j, d, k, p, l, w, h}

Benchmarking with document classification:

K	Algorithm	AG News	Yahoo!	Wiki	Book	CMU
50	Random	48.3%	27.4%	56.9%	31.0%	67.4%
50	Spectral	66.0%	39.8%	71.1%	44.4%	69.5%
50	Improved	**68.5%**	**40.1%**	**71.5%**	**44.7%**	**71.8%**
100	Random	55.5%	33.3%	68.4%	30.0%	67.4%
100	Spectral	72.7%	47.2%	**81.6%**	45.2%	70.0%
100	Improved	**76.8%**	**49.0%**	80.1%	**46.3%**	**70.7%**
200	Random	64.0%	41.7%	80.8%	28.2%	66.8%
200	Spectral	78.2%	51.7%	85.6%	**44.4%**	68.7%
200	Improved	**80.7%**	**54.7%**	**86.5%**	43.4%	**69.0%**
400	Random	72.8%	49.4%	87.8%	28.9%	66.8%
400	Spectral	81.5%	56.3%	88.0%	42.1%	67.9%
400	Improved	**83.1%**	**58.6%**	**89.0%**	**44.4%**	**68.4%**

TU/e

How to evaluate?

Question. Is the block Markov model a good model for the data? Is it worthwhile to research different models?

Answer. Good means better than other models. Let’s compare.

Comparison-based evaluation methods.

1. Benchmarking to compare alternative algorithms.
2. Log-likelihood ratio to compare to alternative models.

Question. Comparison-based methods require an alternative. What if I don’t have prior knowledge of an alternative?

Answer. Do a hypothesis test for your model.

Problem. Hypothesis testing complex data for a simple model always rejects. This is not informative.

Question. How to test only the part of the data which is relevant for clustering?

Answer. The algorithm uses singular value decomposition of $\widehat{N}$.

Let’s look at singular values!

TU/e

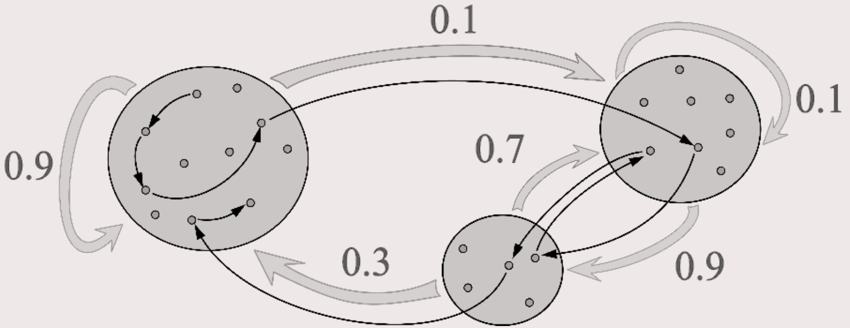
Spectral fingerprints

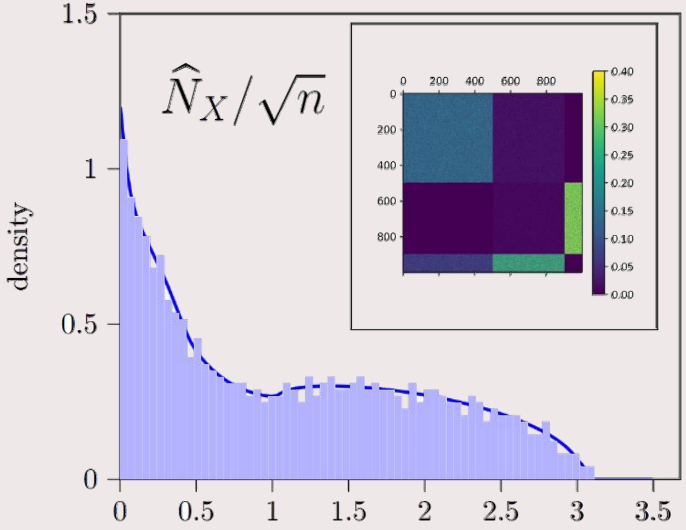
$\widehat{N}$ has many singular values.

We can make a histogram.

Theorem. (J. Sanders, A. Van Werde, 2022)
Assume that $\ell = \Theta(n^2)$. Then, the histogram of singular values of $\widehat{N}/\sqrt{n}$ has a limit as $n \to \infty$.

The limit can be computed in terms of the parameters of the block Markov chain.



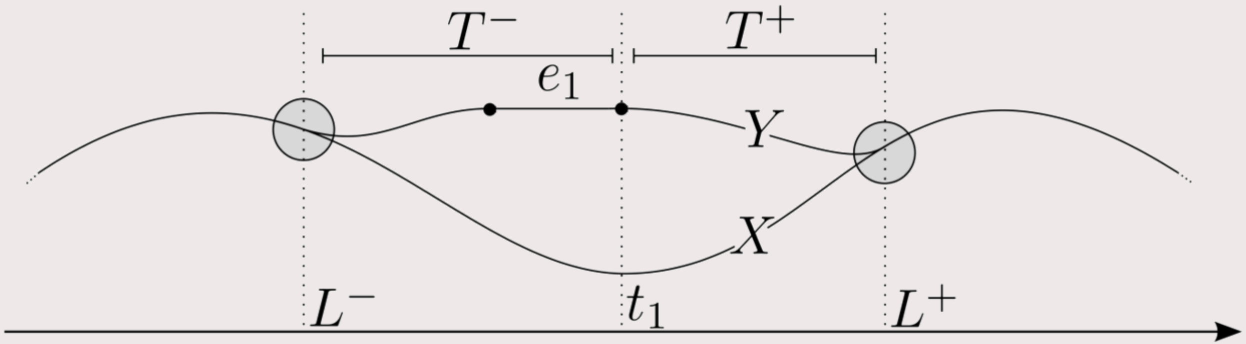


Proof strategy: Coupling

Main difficulty: Markov chains are **dependent** processes.

Coupling: pretend as if the observations X_t are independent up to some error term.

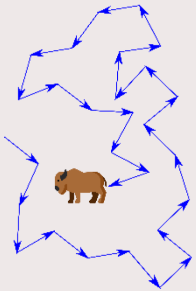
More precisely, given t_1, one constructs a path Y with Y independent of X_{t_1} and $Y_t = X_t$ for most times $t \neq t_1$.

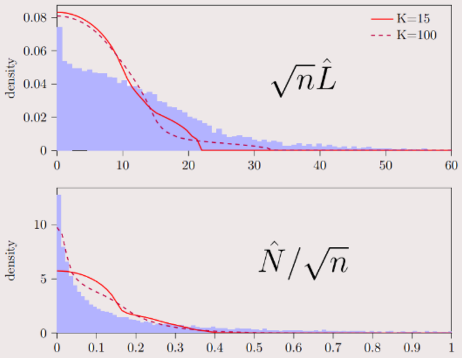


See *"Singular value distribution of dense random matrices with block Markovian dependence."* by Jaron Sanders and Alexander Van Werde at arXiv:2204.13534 for more details.

Spectral fingerprints

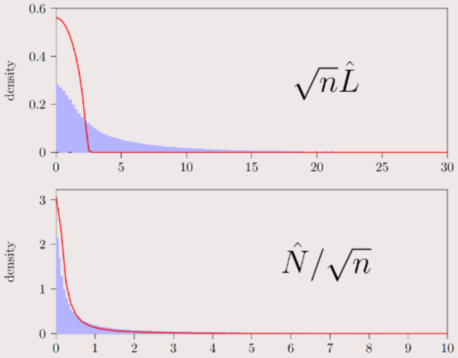
Animal movement



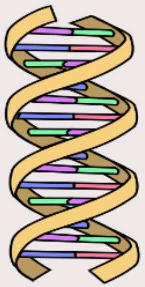


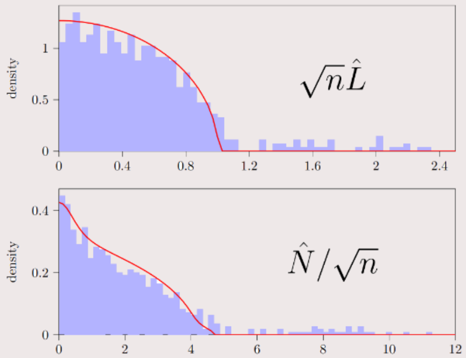
Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas in turpis sit amet risus ultricies tincidunt vitae accumsan enim. Pellentesque ac efficitur enim. Praesent dui magna, mattis at bibendum eu, fringilla eu neque. Nulla vel accumsan arcu. Quisque sagittis, massa quis maximus vestibulum, sapien ligula venenatis ex, eu elementum diam ante a neque. Nunc egestas malesuada interdum. Morbi eget sem in neque pretium vehicula. Duis nulla turpis, efficitur sed lacus at, auctor tempor lorem. Maecenas pharetra et erat non viverra. Morbi pellentesque velit nec risus imperdiet, ac fermentum lectus aliquam. Maecenas rutrum nisi ut turpis ultricies, id placerat quam fringilla. Donec non interdum urna. Nam justo mi, malesuada eu sollicitudin vitae, euismod vitae tortor. Donec feugiat ligula sed sem luctus consectetur id non lectus. Donec ac neque sed metus facilisis rhoncus. Pellentesque iaculis, urna et consequat venenatis, nulla felis laoreet metus, vitae porta risus neque non odio. Sed venenatis mauris magna, sed tincidunt enim suscipit vitae. Donec a commodo ipsum. Proin euismod lacus ac metus finibus sagittis. Duis congue lorem quis velit tempus tincidunt. Ut ultrices rhoncus ipsum, eget aliquam ex dapibus et. Mauris porta, quam sit amet tincidunt aliquet, ex ante tincidunt nibh, et ultricies mi diam ac odio. Duis viverra velit ut dolor ornare ultricies. Phasellus non interdum velit, ac egestas ligula. Cras magna odio, vestibulum in ex ut, scelerisque dapibus tellus. Curabitur facilisis nisl quis erat lobortis efficitur. Donec at ullamcorper nulla, ut eleifend purus. Donec euismod odio quis dui ultricies efficitur. Quisque lobortis lectus tortor, in interdum lorem tempor dictum. Curabitur vitae libero sed ex elementum fringilla. Duis vel sagittis nisl…

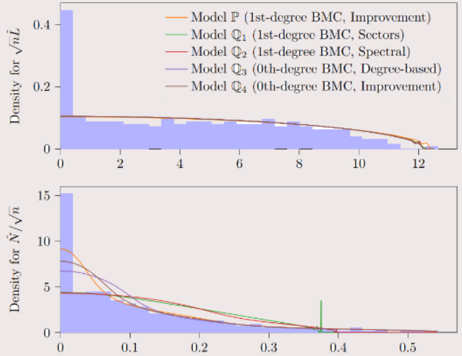


DNA





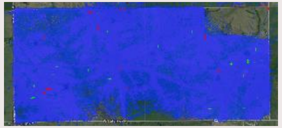
Stock market



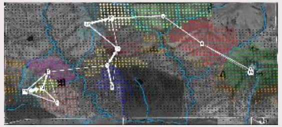
TU/e

Summary

1. Clustering in real-world sequential data can produce insights



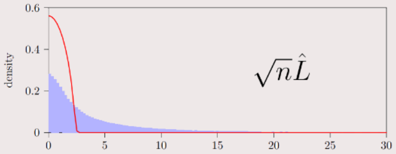
&



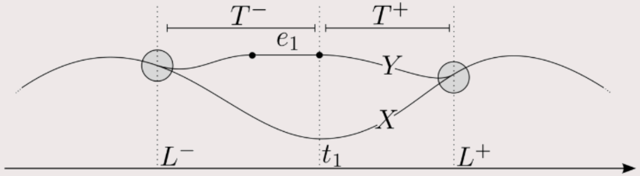
2. Model evaluation is non-trivial. Benchmarking is not sufficient.

K	Algorithm	AG News	Yahoo!	Wiki	Book	CMU
50	Random	48.3%	27.4%	56.9%	31.0%	67.4%
50	Spectral	66.0%	39.8%	71.1%	44.4%	69.5%
50	Improved	**68.5%**	**40.1%**	**71.5%**	**44.7%**	**71.8%**
100	Random	55.5%	33.3%	68.4%	30.0%	67.1%
100	Spectral	72.7%	47.2%	**81.6%**	45.2%	70.0%
100	Improved	**76.8%**	**49.0%**	80.1%	**46.3%**	**70.7%**
200	Random	64.0%	41.7%	80.8%	28.2%	66.8%
200	Spectral	78.2%	51.7%	85.6%	**44.4%**	68.7%
200	Improved	**80.7%**	**54.7%**	**86.5%**	43.4%	**69.0%**
400	Random	72.8%	49.4%	87.8%	28.9%	66.8%
400	Spectral	81.5%	56.3%	88.0%	42.1%	67.9%
400	Improved	**83.1%**	**58.6%**	**89.0%**	**44.1%**	**68.4%**

&



3. Proof technique for dependent data: coupling



Thank you!

Key references and related work:

1. Real-world data analysis:

   Alexander Van Werde, Albert Senen-Cerda, Gianluca Kosmella, and Jaron Sanders. *"Detection and Evaluation of Clusters within Sequential Data."* arXiv:2210.01679, (2022).

2. Spectral fingerprints:

   Jaron Sanders, and **Alexander Van Werde**. *"Singular value distribution of dense random matrices with block Markovian dependence."* arXiv:2204.13534, (2022).

3. Algorithms for clustering in Markov chains:

   Jaron Sanders, Alexandre Proutière, and Se-Young Yun. *"Clustering in block Markov chains."* The Annals of Statistics, (2020).

   Anru Zhang, and Mengdi Wang. *"Spectral state compression of Markov processes."* IEEE Transactions on Information Theory, (2019).

4. Different coupling arguments with random matrices; blocking procedures:

   Banna, Marwa, Florence Merlevède, and Magda Peligrad. *"On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries."* Stochastic Processes and their Applications, (2015)

5. Clustering in reinforcement learning:

   O.-A. Maillard and S. Mannor, *“Latent bandits.,”* in International Conference on Machine Learning, (2014).

   Jedra, Y., Lee, J., Proutière, A., and Yun, S.-Y. *“Nearly Optimal Latent State Decoding in Block MDPs.”* arXiv preprint arXiv:2208.08480, (2022).

6. Community detection in graphs: Stochastic block model

   Holland, P.W., Laskey, K.B. and Leinhardt, S., *“Stochastic blockmodels: First steps.”* Social networks, (1983).

   E. Abbe, *“Community Detection and Stochastic Block Models: Recent Developments,”* Journal of Machine Learning Research, (2017).

An efficient implementation of our algorithms is available: *pip install BMCToolkit*.
https://pypi.org/project/BMCToolkit/