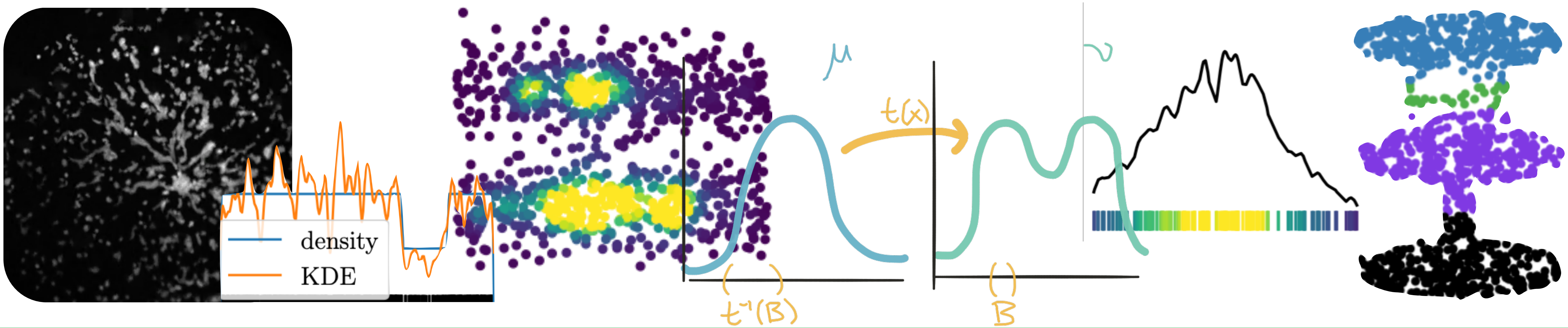


Graph Clustering Dynamics:

From Spectral to Mean Shift via Fokker-Planck

Katy Craig
University of California, Santa Barbara

joint with Nicolás García Trillos (Wisconsin) and Dejan Slepčev (CMU)

University of Minnesota, Institute for Mathematics and its Applications
Data Science Seminar
February 15, 2022

Hellman Fellows Fund

Plan

- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

Plan

- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

Fokker Planck equation

$\rho : [0,T] \to \mathscr{P}(\mathbb{R}^d)$ is a solution of the Fokker-Planck equation if

$$(FP)\begin{cases} \partial_t \rho = \Delta\rho + \mathrm{div}(\rho \nabla V) \\ \rho(0) = \rho_0 \end{cases} \qquad V : \mathbb{R}^d \to \mathbb{R}$$

$d\rho(x) = \rho(x)dx$

Microscopic perspective: $dX_t = \sqrt{2}dB_t - \nabla V(X_t)dt$

Steady state: $Ce^{-V(x)}$

Gradient flow structure: $\partial_t \rho = -\nabla_{W_2}\mathscr{E}(\rho)$, $\mathscr{E}(\rho) = \int \rho \log \rho + \int V\rho$

Motivation for Fokker-Planck equation on a graph:

- Clustering
- Sampling
- Numerical analysis

$$\partial_t \rho = (1-\beta)\Delta\rho + \beta \mathrm{div}(\rho \nabla V)$$

Plan

- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

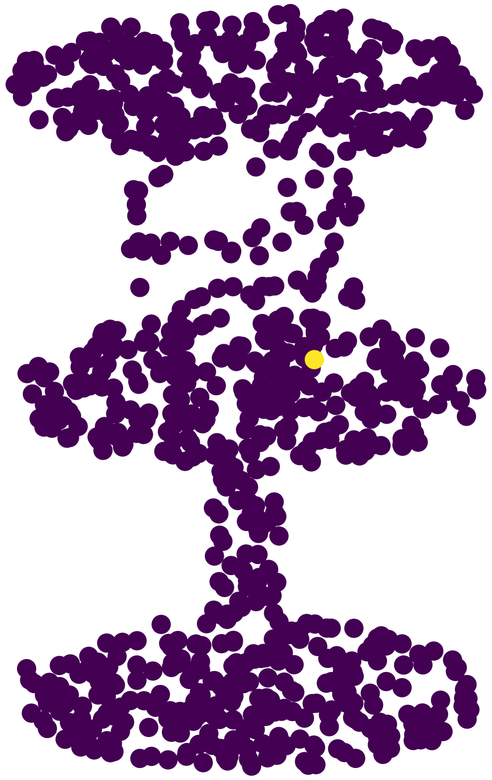
Clustering

Data set $\mathscr{X} = \{x^1, \ldots, x^n\}$

Density	**Geometry**
"clusters" are regions of high concentrations of points, separated by areas of low density	"clusters" are connected regions, separated by bottlenecks
mean shift [Carreira-Perpiñán '16]	spectral clustering [Luxburg '07]

1) Embedding step: $\Psi : \mathscr{X} \to \mathscr{Y}$

2) "Simple" clustering step, e.g., *k*-means



Mean Shift Clustering

$\mathcal{X} = \{x_0^1, \ldots, x_0^n\} \subseteq \mathbb{R}^d$

Given $\hat{q}$, the mean shift algorithm evolves x_0^i via gradient ascent of $\log(\hat{q})$.

kernel density estimate:

$$\hat{q}(x) = \frac{1}{n}\sum_{i=1}^{n} \eta_\delta(|x - x^i|), \quad \eta_\delta(x) = \frac{1}{\delta^d}\eta\left(\frac{x}{\delta}\right), \ \eta \geq 0, \quad \int \eta = 1, \ \ \eta(x) = \eta(|x|)$$

gradient ascent:

$$\begin{cases} x^i(t+1) = x^i(t) + \nabla \log(\hat{q}(x^i(t)) \\ x^i(0) = x^i \end{cases} \qquad (MS)\begin{cases} \frac{d}{dt}x^i(t) = \nabla \log(\hat{q}(x^i(t)) \\ x^i(0) = x_0^i \end{cases}$$

$$\Psi(x_0^i) = x^i(T), \quad T > 0$$

PDE Perspective: $x^i(t)$ solves $(MS) \iff \rho^N(t)$ solves $\rho(x,0) = \delta_{x_0^i}$ and $\partial_t \rho = \nabla \cdot (\rho \nabla V)$ for $V = -\log(\hat{q})$.

Spectral Clustering - Diffusion Maps

Graph Calculus

$\mathcal{X} = \{x_1, \ldots, x_n\}$, $w : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ symmetric

$\mathcal{G} = (\mathcal{X}, w)$ connected

For $\phi : \mathcal{X} \to \mathbb{R}$, define $\nabla_{\mathcal{G}}\phi(x, x') = \phi(x') - \phi(x)$.

For $v : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define $\mathrm{div}_{\mathcal{G}} v(x) = \dfrac{1}{2} \displaystyle\sum_{x'} (v(x, x') - v(x', x)) w(x, x')$.

<u>Definition:</u> The *unnormalized Laplacian* is the operator $\Delta_{\mathcal{G}} = \mathrm{div}_{\mathcal{G}} \circ \nabla_{\mathcal{G}}$.

$$\Delta_{\mathcal{G}} = D - W,\ W_{ij} = w(x^i, x^j),\ D = \mathrm{diag}(d^1, \ldots, d^n),\ d^i = \Sigma_{j \neq i} w(x^i, x^j)$$

<u>Definition:</u> The *Coifman-Lafon Laplacian* is the operator $L_\alpha^{rw} = I - D_\alpha^{-1} W_\alpha$,

$W_\alpha = D^{-\alpha} W D^{-\alpha}$ and $D_\alpha = \mathrm{diag}(d_\alpha^1, \ldots, d_\alpha^n)$, $d_\alpha^i = \displaystyle\sum_{j \neq i} (W_\alpha)_{ij}$

Spectral Clustering - Diffusion Maps

[Coifman Lafon '06]

There exists an orthonormal wrt. $\langle D_\alpha \cdot , \cdot \rangle$ basis of left e-vectors $\{\phi_1, \ldots, \phi_k\}$, corresponding to the first k nonzero e-values of L_α^{rw}.

$$\Psi(x^i) = \begin{bmatrix} \lambda_1^m \phi_1(x^i) \\ \vdots \\ \lambda_k^m \phi_k(x^i) \end{bmatrix}, \quad m \in \mathbb{N}$$

Dynamic interpretation: $-L_\alpha^{rw}$ is a *transition rate matrix*

Definition: $Q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *transition rate matrix* if

1. $Q(x, y) \geq 0$ for $x \neq y$ and 2. $\sum_{y \in \mathcal{X}} Q(x, y) = 0$ for all $x \in \mathcal{X}$.

Diffusion Maps: Continuous Time

$$\mathscr{P}(X) = \left\{ \rho = \sum_{x \in \mathscr{X}} \rho(x)\delta_x : \ \rho : \mathscr{X} \to [0, +\infty) \text{ satisfies } \sum_{x \in \mathscr{X}} \rho(x) = 1 \right\}$$

<u>Definition:</u> A *cts time Markov chain* $\rho : [0,T] \to \mathscr{P}(\mathscr{X})$ is a solution to

$$\begin{cases} \partial_t \rho(y,t) = \sum_{x \in \mathscr{X}} \rho(x,t) Q(x,y) \\ \rho_0(x) = \mu(x) \end{cases} \Longleftrightarrow \begin{cases} \partial_t \rho_t = \rho_t Q \\ \rho_0 = \mu \end{cases} \Longleftrightarrow \rho_t = \mu e^{tQ}$$

Dynamic embedding like mean shift!

$$\Psi(x_i) = \begin{bmatrix} \phi_1 e^{TQ}(x_i) \\ \cdots \\ \phi_k e^{TQ}(x_i) \end{bmatrix}$$

change of basis

$$\Psi(x_i) = \delta_{x_i} e^{TQ}$$
$$= \sum_{l=1}^{n} e^{-T\lambda_l} \frac{\phi_l(x_i)}{d_\alpha(x^i))^{1/2}} \phi_l(x)$$

Diffusion Maps: Continuous Space

<u>Continuum limit:</u>

- $\{x_i\}_{i=1}^n$ iid samples of q
- $w(x,y) = \eta_\epsilon(|x-y|) > 0$
- $Q = -L_\alpha^{rw}/C_{rw}$ for $C_{rw} = M_2(\eta)\epsilon^2/M_0(\eta)$,

As $q_n := \Sigma_{i=1}^n \delta_{x^i} \to q$ and $\epsilon \to 0$ slowly,

$$\rho Q \xrightarrow{n\to+\infty} \Delta_{\mathscr{M}}\rho - 2(1-\alpha)\mathrm{div}_{\mathscr{M}}(\rho \nabla_{\mathscr{M}} \log(q))$$

[Coifman Lafon '06], [Singer'06], [García Trillos Slepcev'18], [Calder, García Trillos '19], [Cheng, Wu '20],…

$$\partial_t \rho_t = \rho_t Q \xrightarrow{n\to+\infty} \partial_t \rho = \Delta_{\mathscr{M}}\rho - 2(1-\alpha)\mathrm{div}_{\mathscr{M}}(\rho \nabla_{\mathscr{M}} \log(q))$$

Diffusion Maps: Cts Time and Space

$\partial_t \rho = \Delta_{\mathscr{M}} \rho - 2(1-\alpha)\mathrm{div}_{\mathscr{M}}(\rho \nabla_{\mathscr{M}} \log(q))$

$\alpha = 1$: Laplace-Beltrami operator, no density, pure geometry

$\alpha = 1/2$: Fokker-Planck equation

$\alpha = 0$: normalized graph laplacian, "maximal density"

After a change of variables, $\tilde{\rho}(x,t) = \rho(x,(3-2\alpha)t)$, $\beta_\alpha = (2-2\alpha)/(3-2\alpha)$

$\partial_t \rho = (1-\beta_\alpha)\Delta_{\mathscr{M}} \rho + \beta_\alpha \mathrm{div}_{\mathscr{M}}(\rho \nabla V), \quad V = -\nabla_{\mathscr{M}} \log(q)$

A Fokker-Planck equation on graphs!

But…

- fixed choice of external potential $V = -\log(q)$, at both discrete & ctm
- degenerates as $\alpha \to -\infty$

Goal

- How can we use the dynamic perspective of diffusion maps to define a true Fokker-Planck equation on a graph, for general external potentials?

- What is the clustering behavior?

Plan

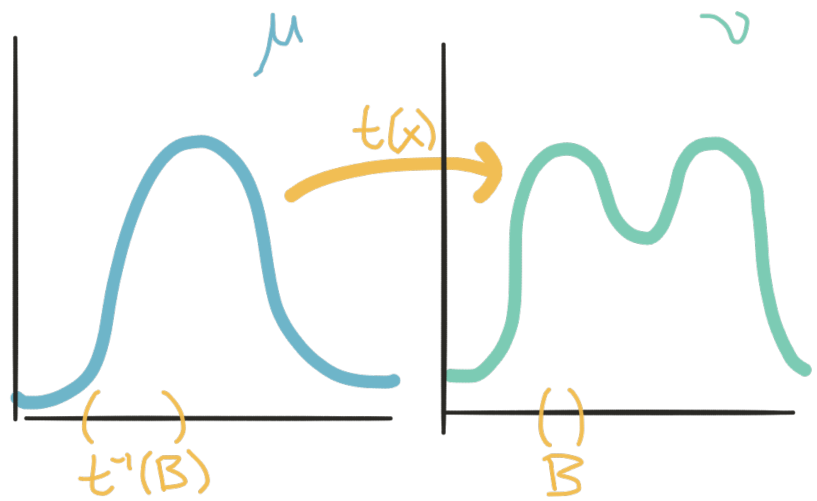
- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

Wasserstein metric

The *Wasserstein distance* between $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ is

$$W_2(\mu,\nu) := \inf\left\{ \underbrace{\left(\int |t(x)-x|^2 d\mu(x)\right)^{1/2}}_{\text{effort to rearrange μ to look like ν, using t(x)}} : \underbrace{t\#\mu = \nu}_{\text{t sends μ to ν}} \right\}$$

where $t\#\mu = \nu$ if $\nu(B) = \mu(t^{-1}(B))$



Alternatively [Benamou, Brenier '00],

$$W_2^2(\mu_0,\mu_1) = \inf\left\{ \int_0^1 \int_{\mathbb{R}^d} |v(x,t)|^2 d\mu(x,t)dt : \partial_t \mu + \nabla(\mu v) = 0 \right\}$$

Gradient flows $\partial_t \rho(t) = -\nabla_{W_2} E(\rho(t))$

Examples:

energy functional	Wasserstein gradient flow
$E(\rho) = \int \rho \log \rho$	$\frac{d}{dt}\rho = \Delta \rho$
$E(\rho) = \frac{1}{m-1} \int \rho^m$	$\frac{d}{dt}\rho = \Delta \rho^m$
$E(\rho) = \int V \rho$	$\frac{d}{dt}\rho = \nabla \cdot (\nabla V \rho)$
$E(\rho) = \int (K * \rho)\rho$	$\frac{d}{dt}\rho = \nabla \cdot (\nabla (K * \rho)\rho)$
$E(\rho) = \int V \rho + \int \rho \log \rho$	$\frac{d}{dt}\rho = \Delta \rho + \nabla \cdot (\nabla V \rho)$

$$\partial_t \rho + \nabla \cdot (\rho v[\rho]) = 0, \quad v[\rho] = -\nabla_{W_2} E(\rho) = -\nabla \frac{\partial E}{\partial \rho}$$

Plan

- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

Wasserstein metric(s) on graphs

Graph continuity equation

$$\rho = \sum_{x \in X} \rho(x)\delta_x \in \mathscr{P}(\mathscr{X}) \text{ , } v : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$$

$\partial_t \rho + \mathrm{div}_{\mathscr{G}}(\bar{\rho} v) = 0$ for $\bar{\rho} : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ interpolating ρ on the edges

Graph action

$$\int_0^t \sum_{x,y \in \mathscr{G}} |v_t(x,y)|^2 w(x,y) d\rho_t(x) dt$$

How to define the interpolating function $\bar{\rho}$?

Choices of density interpolation

arithmetic: $\bar{\rho}(x,y) = \dfrac{\rho(x)+\rho(y)}{2}$

induces a true metric, but GFs not positivity preserving
[Chow, Li, Zhou '18]

logarithmic: $\bar{\rho}(x,y) = \dfrac{\rho(x)-\rho(y)}{\log(\rho(x))-\log(\rho(y))}$

induces a true metric, but support of GF can't expand
[Maas '11], [Mielke '11], [Gigli, Maas '13]

upwinding: $\bar{\rho}(x,y) = \begin{cases} \rho(x) & \text{if } v(x,y) \geq 0, \\ \rho(y) & \text{if } v(x,y) < 0. \end{cases}$

preserves positivity, support can expand, but quasi metric and diff. nonlinear
[Chow, Huang, Li, Zhou '12], [Chen, Georgiou, Tannenbaum '18]
[Esposito, Patacchini, Schlichting, Slepčev '21]

Graph GF: drift

Energy: $\mathscr{V}(\rho) = \sum_{x \in \mathscr{X}} V(x)\rho(x)$

Gradient Flow:

$$\partial_t \rho_t(y) = \sum_{x \in \mathscr{X}} \rho_t(x) Q_V(x,y), \quad Q_V(x,y) := \begin{cases} ((V(x) - V(y))_+ w(x,y) & \text{for } x \neq y, \\ -\sum_{z \neq x} (V(x) - V(z))_+ w(x,y) & \text{for } x = y\,. \end{cases}$$

<u>Formal Theorem</u> [C., García-Trillos, Slepčev '21]:

- $\{x_i\}_{i=1}^n$ iid samples of q
- $w(x,y) = \eta_\epsilon(|x-y|) > 0$
- $Q = Q_V / C_{MS}$ for $C_{MS} = 2M_2(\eta) d n \epsilon^2$.

As $q_n := \Sigma_{i=1}^n \delta_{x^i} \to q$ and $\epsilon \to 0$ slowly

$$\rho Q \xrightarrow{n \to +\infty} \operatorname{div}_{\mathscr{M}}(\rho q \nabla_{\mathscr{M}} V).$$

See also [Esposito, Patacchini, Schlichting, Slepčev '21] for $n \to +\infty, \epsilon > 0$.

Graph GF: drift

$$\partial_t \rho + \mathrm{div}_{\mathcal{M}}(\rho q \nabla_{\mathcal{M}} V) = 0$$

When $V = \log(q)$, this is not quite mean shift.

A Wasserstein gradient flow with nontrivial mobility, $h(\mu(x)) = \mu(x)q(x)$:

$$W_{2,h}^2(\mu_0, \mu_1) = \inf \left\{ \int_0^1 \int_{\mathbb{R}^d} |v(x,t)|^2 h(\mu(x,t), x) dx dt : \partial_t \mu + \nabla(h(\mu)v) = 0 \right\}$$

[Dolbeault, Nazaret, Savaré '08]

Modifying the ground metric on the underlying space $\mathbb{R}^d$:

$$d_q(x,y) = \inf \left\{ \int_0^1 \sqrt{q(\gamma(t))^{-1}} |\dot{\gamma}(t)| dt : \gamma \in AC([0,1]; \mathbb{R}^d, \gamma(0) = x, \gamma(1) = y \right\}$$

[Lisini '09]

$$V(x) = -\frac{1}{q(x)}$$

Fokker-Planck on graphs

GF of potential energy: $\partial_t \rho_t = \rho_t Q_V / C_{MS}$,

$$Q_V(x,y) = \begin{cases} ((V(x) - V(y))_+ w(x,y) & \text{for } x \neq y, \\ -\sum_{z \neq x} (V(x) - V(z))_+ w(x,y) & \text{for } x = y\,. \end{cases}$$

Fokker-Planck: $\partial_t \rho_t = \rho_t Q_\alpha$ **for** $Q_\beta = -(1-\beta) L_1^{rw} / C_{rw} + \beta Q_V / C_{MS}$

- Formal continuum limits:

$$\partial_t \rho = (1-\beta)\Delta_{\mathscr{M}}\rho + \beta \operatorname{div}_{\mathscr{M}}(\rho q \nabla_{\mathscr{M}} V) \text{ for } \alpha = 1$$

- A true Fokker-Planck equation, including both endpoints at all timescales.

- Flexibility in choice of external potential

Clustering Algorithm

Given $q \in \mathscr{P}(\Omega)$, $\Omega \subset\subset \mathbb{R}^d$, let $\{x_i\}_{i=1}^n$ be iid samples from q.

$$w(x,y) = \eta_\epsilon(|x-y|), \quad \eta_\epsilon(x) = e^{-x^2/(2\epsilon^2)}/(2\pi\epsilon^2)^{d/2}$$

$\epsilon = \sqrt{2} \max_i \min_{j:j\neq i} |x_i - x_j|$ in one dimension

$$\hat{q}(x) = -\frac{1}{n}\sum_{y\in\mathscr{X}} \eta_\delta(|x-y|), \quad \delta = \sqrt{2}\left(\frac{|\Omega|}{n}\right)^{1/2}$$

Algorithm 1 Dynamic Clustering Algorithm

Input: $\{x_i\}_{i=1}^n$, ε, δ, t, k,Q

$\hat{\Psi}_Q(x_i) = (e^{tQ})_{(i,j=1,\ldots n)}$ for $i = 1, \ldots, n$

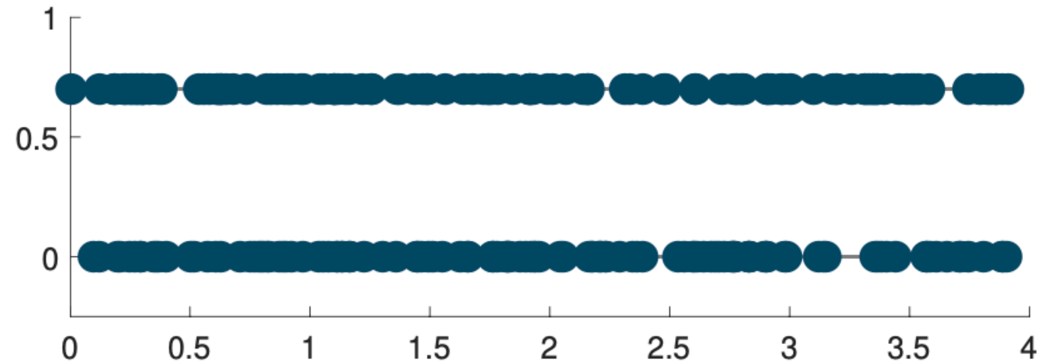
$l_m = \mathrm{Kmeans.fit}(\hat{\Psi}_Q(x_1), \ldots, \hat{\Psi}_Q(x_n))$ with $\mathrm{n_{clusters}} = k$

Plan

- Main goal: Fokker-Planck on a graph
- Motivation: density vs geometry in clustering
- Wasserstein gradient flows
- Wasserstein gradient flows on graphs
- Numerical examples

Numerics: Graph Mean Shift

initial conditions

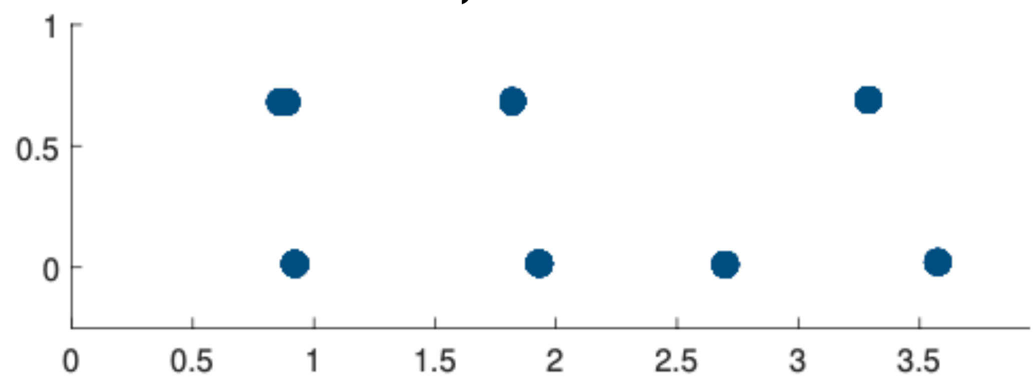


- **restricts dynamics to data**
- **sensitive to δ, noise**

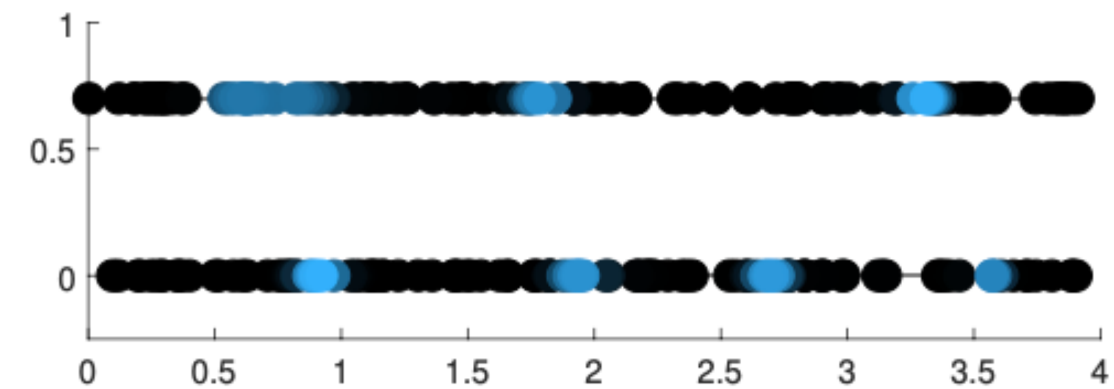
$n = 280$

$\epsilon = 0.3$

long time behavior, $\delta = 0.25$

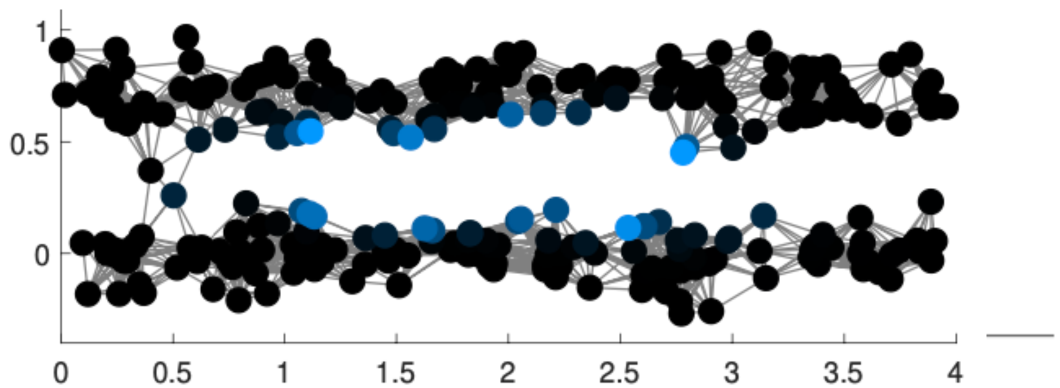


MS

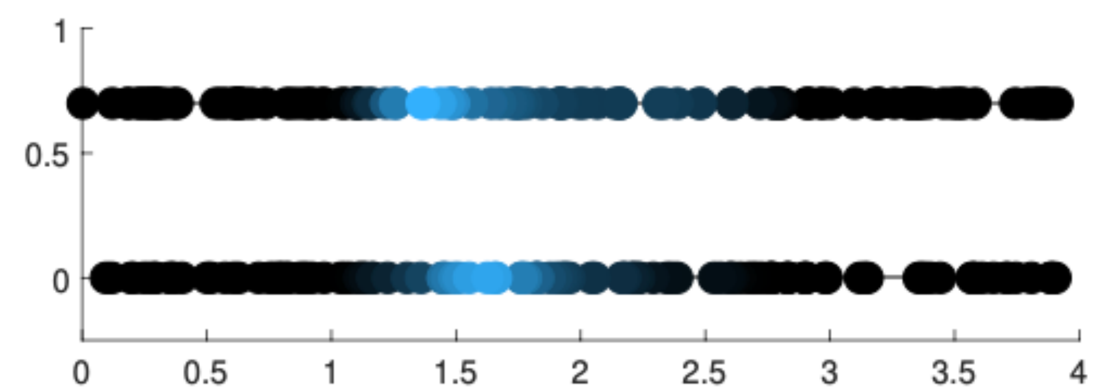


GMS

long time behavior, $\delta = 0.71$

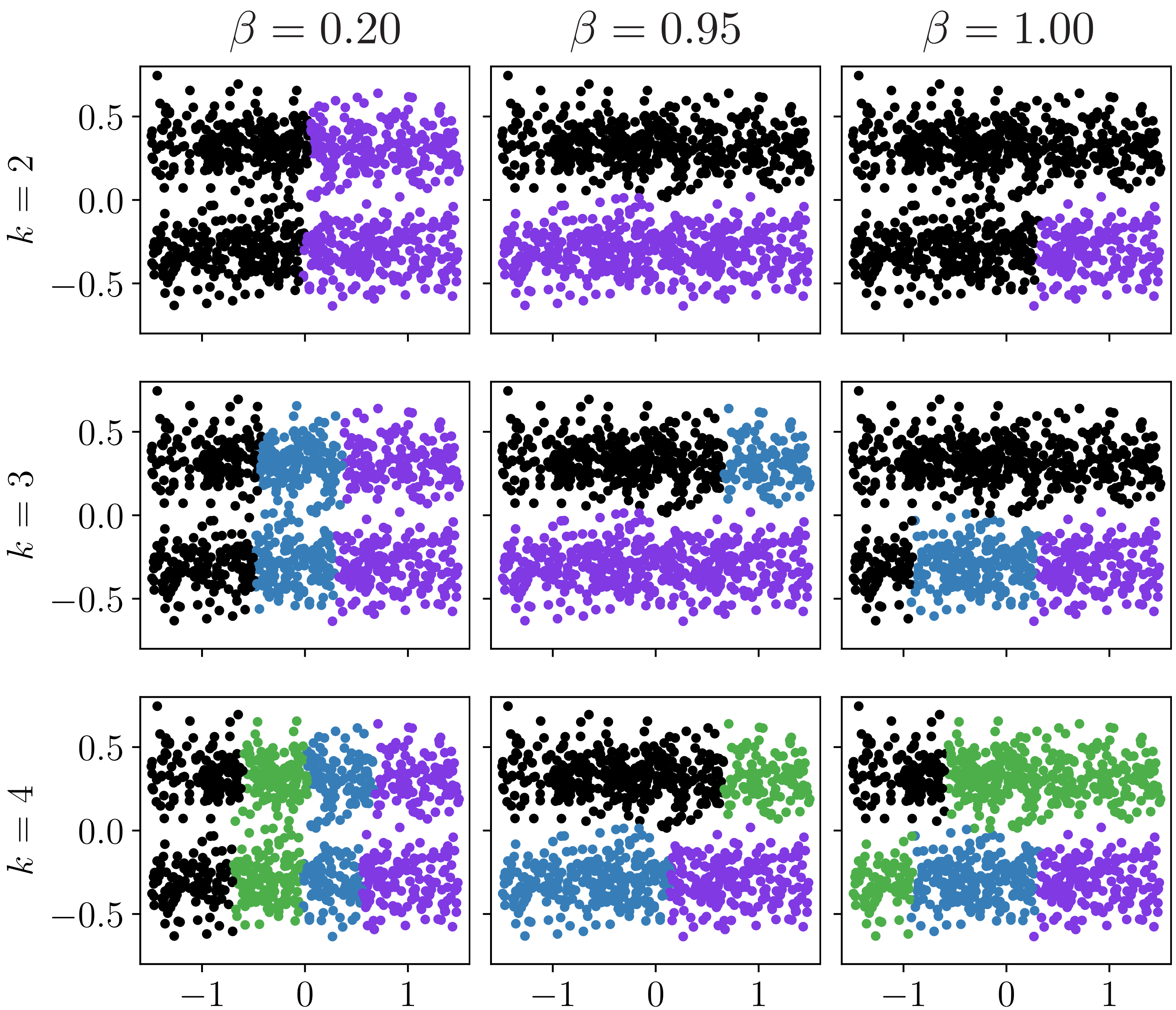


GMS



GMS

Numerics: Graph Fokker-Planck



A small amount of diffusion helps graph mean shift overcome the problems of a noisy KDE and “getting trapped”.

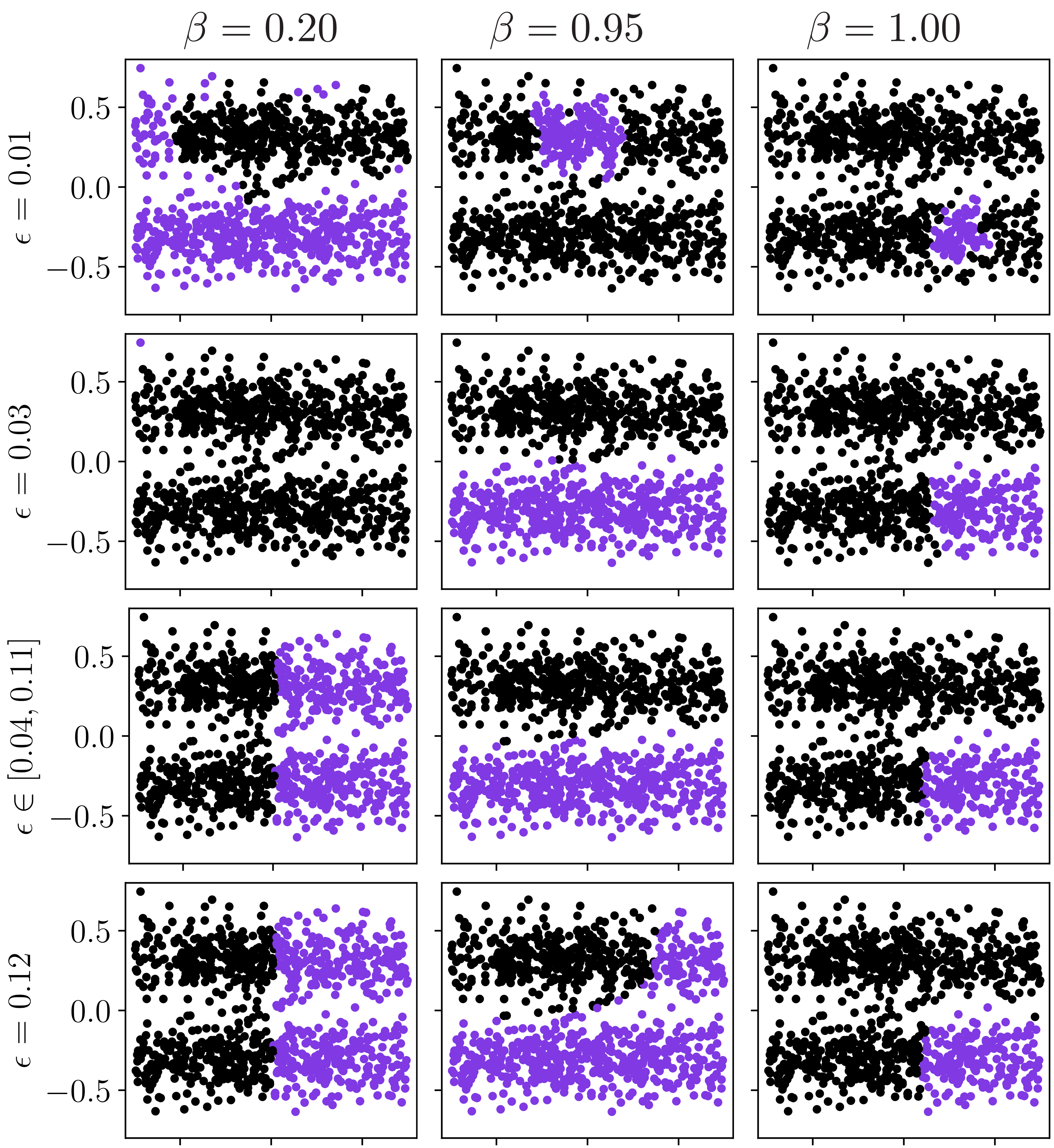
$n = 965$

$\epsilon = 0.04$

$\delta = 0.10$

$T = 10$

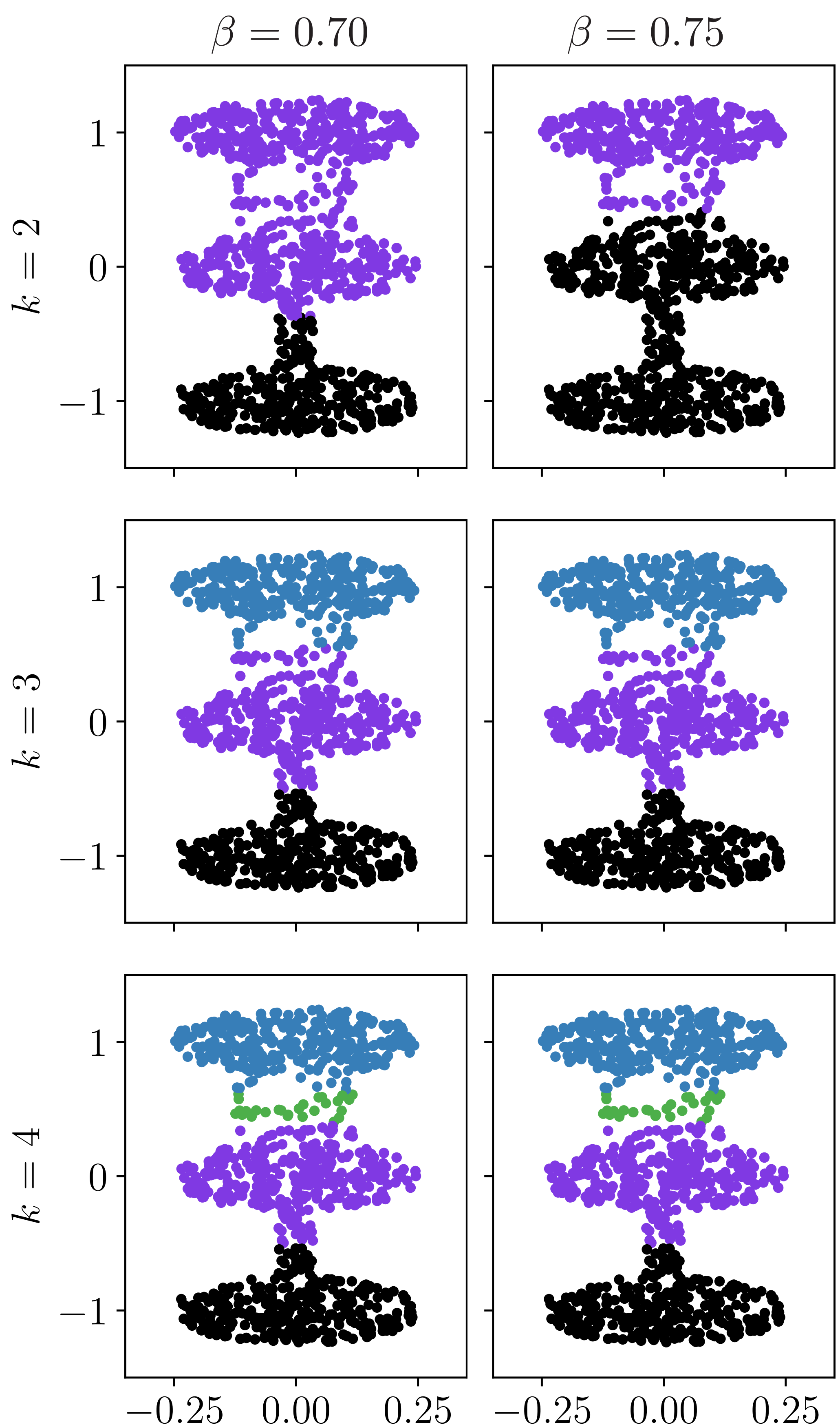
Numerics: Graph Fokker-Planck



- **Decreasing the connectivity parameter ϵ isn't enough to save pure diffusion methods.**
- **Graph Fokker-Planck performs well for a wide range of ϵ.**

$n = 965$
$\delta = 0.10$
$T = 10$

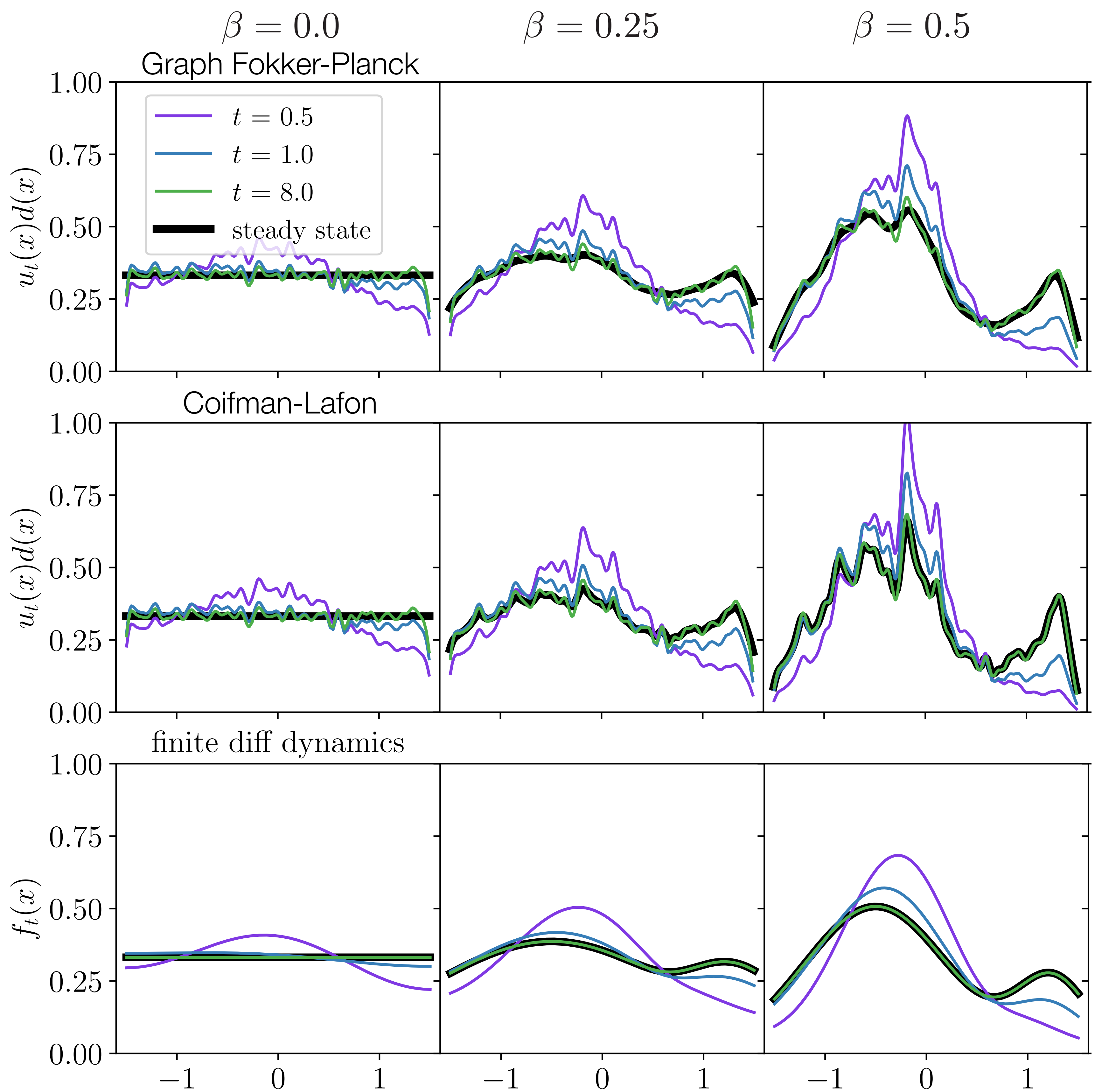
Density vs geometry



Choosing the "right" balance between density and geometry depends on modeling assumptions.

$n = 966$
$\epsilon = 0.07$
$\delta = 0.05$
$T = 10$

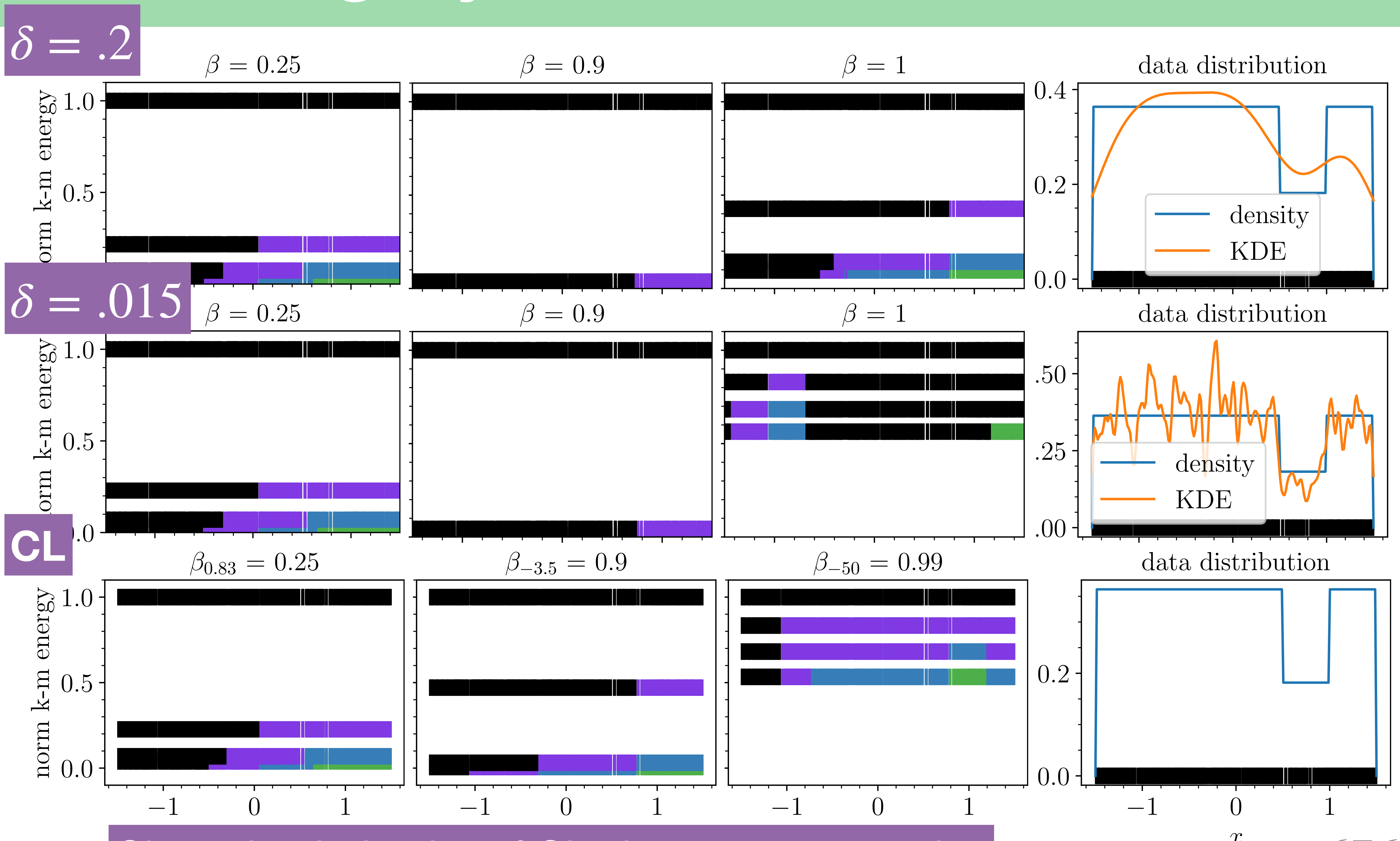
GFP vs Coifman Lafon



- **graph dynamics agree well with continuum PDE**
- **Graph Fokker-Planck steady state depending on KDE bandwidth δ**
- **Coifman-Lafon steady state depending on KDE bandwidth ϵ**

$n = 625$

Clustering Dynamics and KDE



Clustering behavior of CL also appears to rely on density estimator with bandwidth ϵ.

$n = 676$
$T = 30$

Future directions

- How can analysis of eigenvalues lead to appropriate choices of T? Hierarchical clustering method?
- Sampling on graphs? Stochastic particle method?
- Can we combine logarithmic & unwinding interpolation, via inf-convolution or product structure, to get gradient flow structure of graph FP? Rigorous proof of continuum limit?
- Numerical analysis -> data analysis?

Thank you!