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
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

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

\[
(FP) \quad \begin{cases} 
\partial_t \rho = \Delta \rho + \text{div}(\rho \nabla V) & V : \mathbb{R}^d \rightarrow \mathbb{R} \\
\rho(0) = \rho_0 
\end{cases}
\]

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

Steady state: \( C e^{-V(x)} \)

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

Motivation for Fokker-Planck equation on a graph:

- Clustering
- Sampling
- Numerical analysis

\[
\partial_t \rho = (1 - \beta) \Delta \rho + \beta \text{div}(\rho \nabla V)
\]
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 $\mathcal{X} = \{x^1, \ldots, x^n\}$

**Density**

“clusters” are regions of high concentrations of points, separated by areas of low density

mean shift [Carreira-Perpiñán ’16]

**Geometry**

“clusters” are connected regions, separated by bottlenecks

spectral clustering [Luxburg ’07]

1) Embedding step: $\Psi : \mathcal{X} \rightarrow Y$

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:**

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

**gradient ascent:**

\[
\begin{cases}
  x^i(t + 1) = x^i(t) + \nabla \log(\hat{q}(x^i(t))) \\
  x^i(0) = x^i
\end{cases}
\]

\( (MS) \)

\[
\begin{cases}
  \frac{d}{dt} x^i(t) = \nabla \log(\hat{q}(x^i(t))) \\
  x^i(0) = x^i_0
\end{cases}
\]

\( \Psi(x^i_0) = x^i(T), \quad T > 0 \)

**PDE Perspective:** \( x^i(t) \) solves \( (MS) \) \( \iff \) \( \rho^N(t) \) solves \( \rho(x,0) = \delta_{x_0^i} \) and

\[
\partial_t \rho = \nabla \cdot (\rho \nabla V) \text{ for } V = -\log(\hat{q}).
\]
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 \( \text{div}_\mathcal{G} v(x) = \frac{1}{2} \sum_{x'} (v(x, x') - v(x', x))w(x, x') \).

**Definition:** The unnormalized Laplacian is the operator \( \Delta_\mathcal{G} = \text{div}_\mathcal{G} \circ \nabla_\mathcal{G} \).

\[
\Delta_\mathcal{G} = D - W, \quad W_{ij} = w(x^i, x^j), \quad D = \text{diag}(d^1, \ldots, d^n), \quad d^i = \sum_{j \neq i} w(x^i, x^j)
\]

**Definition:** The Coifman-Lafon Laplacian is the operator \( L^r_\alpha = I - D^{-1}_\alpha W_\alpha \),

\[
W_\alpha = D^{-\alpha} W D^{-\alpha} \quad \text{and} \quad D_\alpha = \text{diag}(d^1_\alpha, \ldots, d^n_\alpha), \quad d^i_\alpha = \sum_{j \neq i} (W_\alpha)_{ij}
\]
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 \( L^{rw}_\alpha \).

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

Dynamic interpretation: \(-L^{rw}_\alpha\) is a transition rate matrix

**Definition:** \( Q : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) 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 \( x \in \mathcal{X} \).
Definition: A *continuous time Markov chain* $\rho : [0,T] \to \mathcal{P}(\mathcal{X})$ is a solution to

\[
\begin{align*}
\frac{\partial \rho(y, t)}{\partial t} &= \sum_{x \in \mathcal{X}} \rho(x, t) Q(x, y) \\
\rho(0)(x) &= \mu(x)
\end{align*}
\]

\[
\begin{align*}
\frac{\partial \rho_t}{\partial t} &= \rho_t Q \\
\rho_0 &= \mu \\
\rho_t &= \mu e^{tQ}
\end{align*}
\]

\[
\Psi(x_i) = \begin{bmatrix} 
\phi_1 e^{TQ(x_i)} \\
\vdots \\
\phi_k e^{TQ(x_i)} 
\end{bmatrix}
\]

change of basis

\[
\Psi(x) = \delta_{x_i} e^{TQ} = \sum_{l=1}^{n} e^{-T\lambda_l} \frac{\phi_l(x_i)}{d_{\alpha}(x_i)^{1/2}} \phi_l(x)
\]
**Continuum limit:**

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

As \( q_n := \sum_{i=1}^n \delta_{x_i} \rightarrow q \) and \( \epsilon \rightarrow 0 \) slowly,

\[
\rho Q \xrightarrow{n \rightarrow +\infty} \Delta \mathcal{M} \rho - 2(1 - \alpha) \text{div}_\mathcal{M}(\rho \nabla \mathcal{M} \log(q))
\]

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

\[
\partial_t \rho_t = \rho_t Q \xrightarrow{n \rightarrow +\infty} \partial_t \rho = \Delta \mathcal{M} \rho - 2(1 - \alpha) \text{div}_\mathcal{M}(\rho \nabla \mathcal{M} \log(q))
\]
Diffusion Maps: Cts Time and Space

\[ \partial_t \rho = \Delta_M \rho - 2(1 - \alpha) \text{div}_M(\rho \nabla_M \log(q)) \]

\( \alpha = 1 \): 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, \( \tilde{\rho}(x, t) = \rho(x, (3 - 2\alpha)t), \quad \beta_\alpha = (2 - 2\alpha)/(3 - 2\alpha) \)

\[ \partial_t \rho = (1 - \beta_\alpha) \Delta_M \rho + \beta_\alpha \text{div}_M(\rho \nabla V), \quad V = - \nabla_M \log(q) \]

A Fokker-Planck equation on graphs!

But…

- fixed choice of external potential \( V = - \log(q) \), at both discrete & ctm
- degenerates as \( \alpha \to - \infty \)
• 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 $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is

$$W_2(\mu, \nu) := \inf \left\{ \left( \int |t(x) - x|^2 d\mu(x) \right)^{1/2} : t\#\mu = \nu \right\}$$

effort to rearrange $\mu$ to look like $\nu$, using $t(x)$.

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

Alternatively [Benamou, Brenier ’00],

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

\[ \partial_t \rho(t) = - \nabla_{W_2} E(\rho(t)) \]

**Examples:**

<table>
<thead>
<tr>
<th>Energy functional</th>
<th>Wasserstein gradient flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ E(\rho) = \int \rho \log \rho ]</td>
<td>[ \frac{d}{dt} \rho = \Delta \rho ]</td>
</tr>
<tr>
<td>[ E(\rho) = \frac{1}{m-1} \int \rho^m ]</td>
<td>[ \frac{d}{dt} \rho = \Delta \rho^m ]</td>
</tr>
<tr>
<td>[ E(\rho) = \int V \rho ]</td>
<td>[ \frac{d}{dt} \rho = \nabla \cdot (\nabla V \rho) ]</td>
</tr>
<tr>
<td>[ E(\rho) = \int (K \ast \rho) \rho ]</td>
<td>[ \frac{d}{dt} \rho = \nabla \cdot (\nabla (K \ast \rho) \rho) ]</td>
</tr>
<tr>
<td>[ E(\rho) = \int V \rho + \int \rho \log \rho ]</td>
<td>[ \frac{d}{dt} \rho = \Delta \rho + \nabla \cdot (\nabla V \rho) ]</td>
</tr>
</tbody>
</table>

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

\[ \rho = \sum_{x \in X} \rho(x) \delta_x \in \mathcal{P}(\mathcal{X}), \quad v : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \]

\[ \partial_t \rho + \text{div}_\mathcal{G}(\bar{\rho} v) = 0 \text{ for } \bar{\rho} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \text{ interpolating } \rho \text{ on the edges} \]

Graph action

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

How to define the interpolating function \( \bar{\rho} \)?
Choices of density interpolation

**Arithmetic:** \( \bar{\rho}(x, y) = \frac{\rho(x) + \rho(y)}{2} \)

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

**Logarithmic:** \( \bar{\rho}(x, y) = \frac{\rho(x) - \rho(y)}{\log(\rho(x)) - \log(\rho(y))} \)

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

Gradient Flow:

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

**Formal Theorem** [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 $C_{MS} = 2M_2(\eta)d\epsilon^2$.

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

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

See also [Esposito, Patacchini, Schlichting, Slepčev ’21] for $n \to +\infty$, $\epsilon > 0$. 
Graph GF: drift

\[ \partial_t \rho + \text{div}_\mathcal{M}(\rho q \nabla_\mathcal{M} V) = 0 \]

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

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

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

[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]

\[
V(x) = -\frac{1}{q(x)}
\]
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 $Q_\beta = - (1 - \beta) L_{1r^w} / C_{rw} + \beta Q_V / C_{MS}$

- Formal continuum limits:

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

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

- Flexibility in choice of external potential
Given $q \in \mathcal{P}(\Omega)$, $\Omega \subset \subset \mathbb{R}^d$, let $\{x_i\}_{i=1}^n$ be iid samples from $q$.

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

$$\varepsilon = \sqrt{2} \max_{i} \min_{j:j\neq i} |x_i - x_j| \text{ in one dimension}$$

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

Algorithm 1 Dynamic Clustering Algorithm

**Input:** $\{x_i\}_{i=1}^n, \varepsilon, \delta, t, k, Q$

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

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

$n = 280$
$\epsilon = 0.3$

**Long time behavior, $\delta = 0.25$**

**Long time behavior, $\delta = 0.71$**
A small amount of diffusion helps graph mean shift overcome the problems of a noisy KDE and “getting trapped”.

In Figure 12, we consider the graph dynamics of $Q = 0, \ldots, 75$ on a two dimensional blue sky, with each node colored according to which cluster it belongs. The fourth row shows the normalized means energy for each number of clusters, considered together with other problems of a noisy KDE for $t = 0, \ldots, 125$.

The data density, which we refer to as $\mathcal{P}$, is given by a piecewise constant function that is equal to $\mathcal{P}(x) = 0$ for $|x| > 0.5$. The columns correspond to $k = 0, \ldots, 4$, in order to optimize agreement on the wide rectangle and $\beta = 0.20, 0.95, 1.00$, on the narrow rectangle height one on the three circles of radius $r = 0.5, 0.95, 1.00$.

Finally, in Figure 11, we investigate how the clustering behavior of $\mathcal{P}$ prefers to make shorter cuts even over parts of the domain where data are dense, which is undesirable for the data considered. On the other hand, the pure mean shift suffers, as in other examples, from the tendency to identify spurious local maxima of KDE as clusters. We observe that adding even a small amount of diffusion goes a long way towards correct dense, which is undesirable for the data considered. On the other hand, the pure mean shift suffers, as in other examples, from the tendency to identify spurious local maxima of KDE as clusters. We observe that adding even a small amount of diffusion goes a long way towards correct clustering.

In Figure 12, we choose $n = 965$, $\epsilon = 0.04$, $\delta = 0.10$, $T = 10$. The first three rows show the clustering behavior for $k = 2, 3, 4$.
• Decreasing the connectivity parameter $\epsilon$ isn’t enough to save pure diffusion methods.
• Graph Fokker-Planck performs well for a wide range of $\epsilon$. 

$n = 965$
$\delta = 0.10$
$T = 10$
Choosing the “right” balance between density and geometry depends on modeling assumptions.
• graph dynamics agree well with continuum PDE
• Graph Fokker-Planck steady state depending on KDE bandwidth $\delta$
• Coifman-Lafon steady state depending on KDE bandwidth $\epsilon$

$n = 625$
Clustering Dynamics and KDE

The top row of markers in each frame corresponds to a single cluster (Q), and kernel density estimate. Note that, since no explicit kernel density estimate is used in the construction of clusters (Q), none is shown in the third row. The colors of the samples indicate the clusters to which they belong, and kernel density estimate with bandwidth in the middle row was chosen to match the bandwidth used in the construction of the implicit kernel density estimate which appears to drive the dynamics of diffusion cause the dynamics to ignore the changes in relative density and cluster based on the fairly uniform geometry of the sampling. In the middle row, when the bandwidth in the KDE is small, we still observe good performance for normalization k-m energy and kernel density estimate. On the other hand, when the bandwidth in the KDE is large, we observe good clustering performance for normalization k-m energy and kernel density estimate. The first three columns show clustering with three and four clusters. The middle row shows the dynamics of CL also appears to rely on density estimator with bandwidth c.

The bottom row, the only way to increase the bandwidth of the implicit kernel density estimate would be to increase the graph connectivity parameter. 

Figure 7 illustrates the effect that different choices in data distribution and kernel density estimate. Note that, since no explicit kernel density estimate is used in the construction of clusters (Q), none is shown in the third row. The colors of the samples indicate the clusters to which they belong, and kernel density estimate with bandwidth in the middle row was chosen to match the bandwidth used in the construction of the implicit kernel density estimate which appears to drive the dynamics of diffusion cause the dynamics to ignore the changes in relative density and cluster based on the fairly uniform geometry of the sampling. In the middle row, when the bandwidth in the KDE is small, we still observe good performance for normalization k-m energy and kernel density estimate. On the other hand, when the bandwidth in the KDE is large, we observe good clustering performance for normalization k-m energy and kernel density estimate. The first three columns show clustering with three and four clusters. The middle row shows the dynamics of CL also appears to rely on density estimator with bandwidth c. 

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