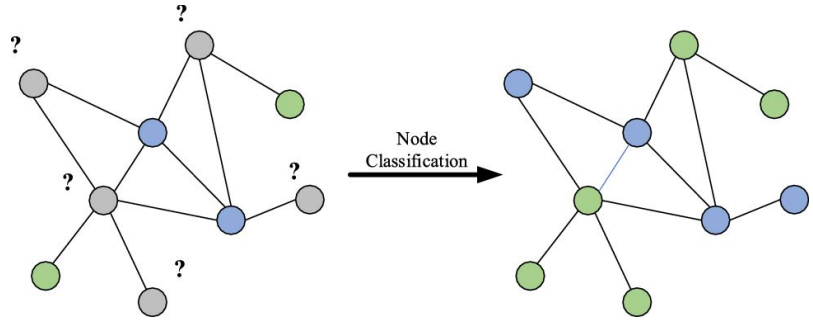
Residual GNNs with Normalization.

Akshay Sakanaveeti

Node classification problem:



- Graph: $G = (V, E)$
- Vertices divided into two types: 0 and 1
- Each vertex has an initial feature vector denoted by $X_v^{(0)} \in \mathbb{R}^2$

Question: How do we classify nodes?

1. **Graph methods (e.g., spectral clustering):** Focus on graph structure but ignore node features
2. **Traditional ML (e.g., logistic regression):** Focus on node features but ignore graph structure
3. **GNNs:** Use *both* features and graph structure

GNNs

1. **Message passing:** Take a **neighborhood average.** Take a **convex combination** with the previous feature

$$X_v^{(t+1)} = (1-\alpha)\, X_v^{(t)} + \frac{\alpha}{\deg(v)} \sum_{u \sim v} W^{(t)} X_u^{(t)}$$

2. **Classification:** $\left(X_v^{(T)} : v \in V\right)$ are then fed to a classification algorithm.

Interpretation:

- $X_v^{(0)}$: feature of the node
- $X_v^{(1)}$: contains information from neighbors
- $X_v^{(2)}$: information from neighbors-of-neighbors

Problem: Oversmoothing

1. We want large T because then $X_v^{(T)}$ contains information of neighbors upto distance T.
2. Issue: Feature become indistinguishable

$$X_v^{(t)} \approx X_u^{(t)} \quad \text{for all } u, v$$

3. When $\alpha \approx 1$, $X_v^{(t)}$ at each step is close to neighborhood average.
4. Only way to mitigate oversmoothing is by taking “alpha” to be small. **In that case, one can approximate the process by a continuous time process.**

Continuous time approximation

When “alpha” is small, one can approximate the process by a continuous time process.

This gives us more insight into rate of collapse, limit point etc.

For example, when $\alpha \approx 0$ and $W^{(t)} = I$, we have

$$X_v^{(t+1)} - X_v^{(t)} = \alpha \left(-X_v^{(t)} + \frac{1}{\deg(v)} \sum_{u \sim v} X_u^{(t)} \right)$$

This is approximated by the continuous-time process:

$$dX_v^{(t)} = -X_v^{(t)} + \frac{1}{\deg(v)} \sum_{u \sim v} X_u^{(t)} \, dt$$

Mitigating oversmoothing

1. Residual GNNs

- Choose i.i.d weight matrices $W^{(t)}$ from some distribution
- Intuition: adding noise slows the collapse
- Effect: oversmoothing is **slowed**, but **not prevented**

2. Normalization After Each Layer

- After update, renormalize features
- Forces all features to lie on a **circle** (unit circle in $\mathbb{R}^2$)
- Used in some transformer-style models

We will focus on what happens when these two methods are combined.

Combining two residual GNNs with normalisation

Step 1 (Update): $$Y_v^{(t+1)} = (1-\alpha)X_v^{(t)} + \frac{\alpha}{\deg(v)} \sum_{u \sim v} W^{(t)} X_u^{(t)}$$

Step 2 (Normalize): $$X_v^{(t+1)} = \frac{Y_v^{(t+1)}}{\|Y_v^{(t+1)}\|}$$

- When $\alpha \approx 1$, oversmoothing is expected
- So we focus on the regime $\alpha \approx 0$

We focus on the case where entries of W^t are iid $N(0, \sigma^2)$

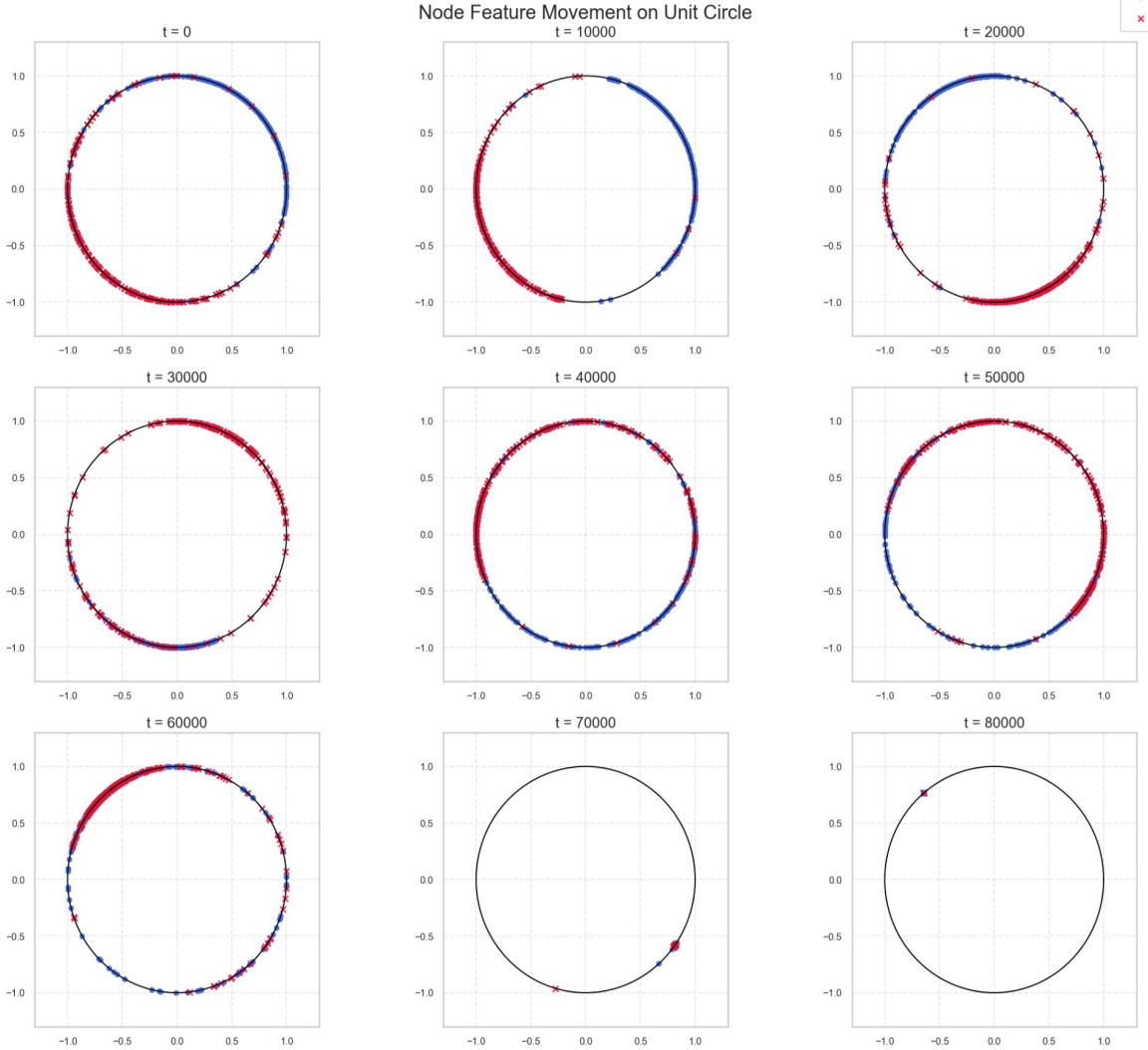
Simulations

1. G= SBM, alpha = 0.001
2. Number of vertices = 500
3. Intra cluster connection probability = 1/25
4. Intertype = 1/100
5. Initial features:

   Type 0- iid N(mu,I_2)

   Type 1- iid N(-mu,I_2)

   (then normalised)
6. W^t have iid N(0,sigma^2) entries

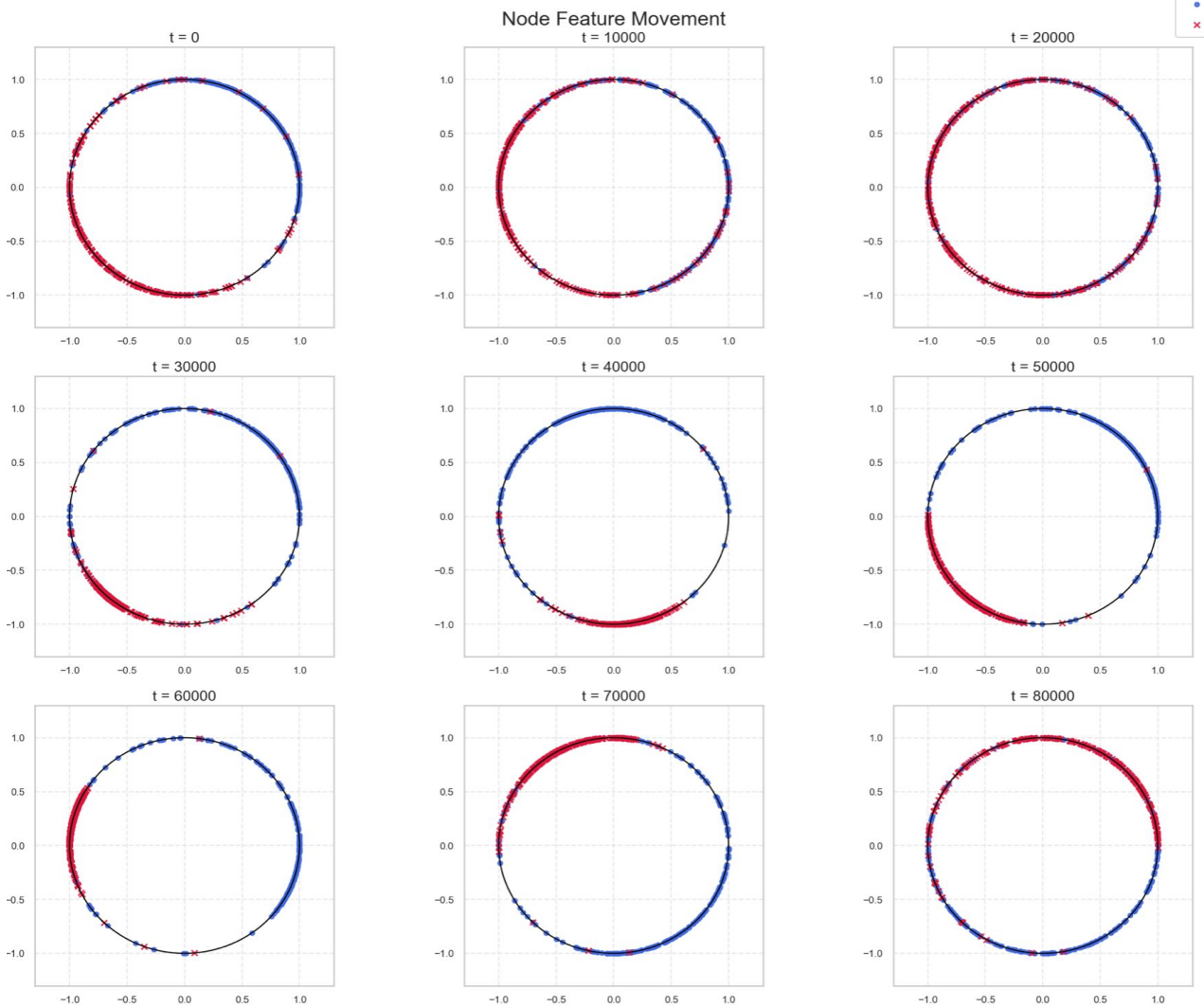


So this model does not

prevent oversmoothing.

Changing σ^2 to σ^2/α

prevents oversmoothing.



Questions

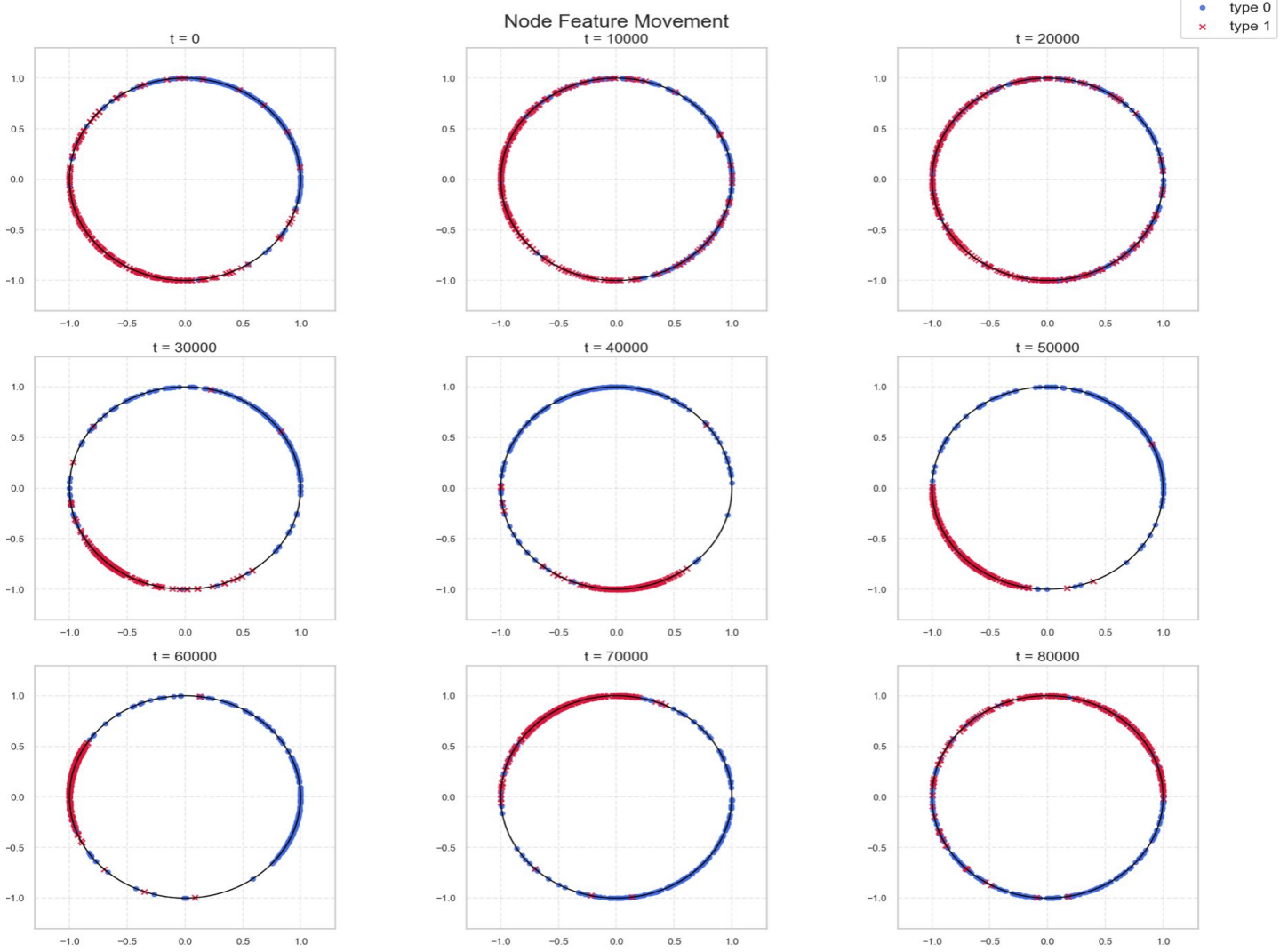
1. So this model prevents all features collapsing to a deterministic point.
2. **Question:**

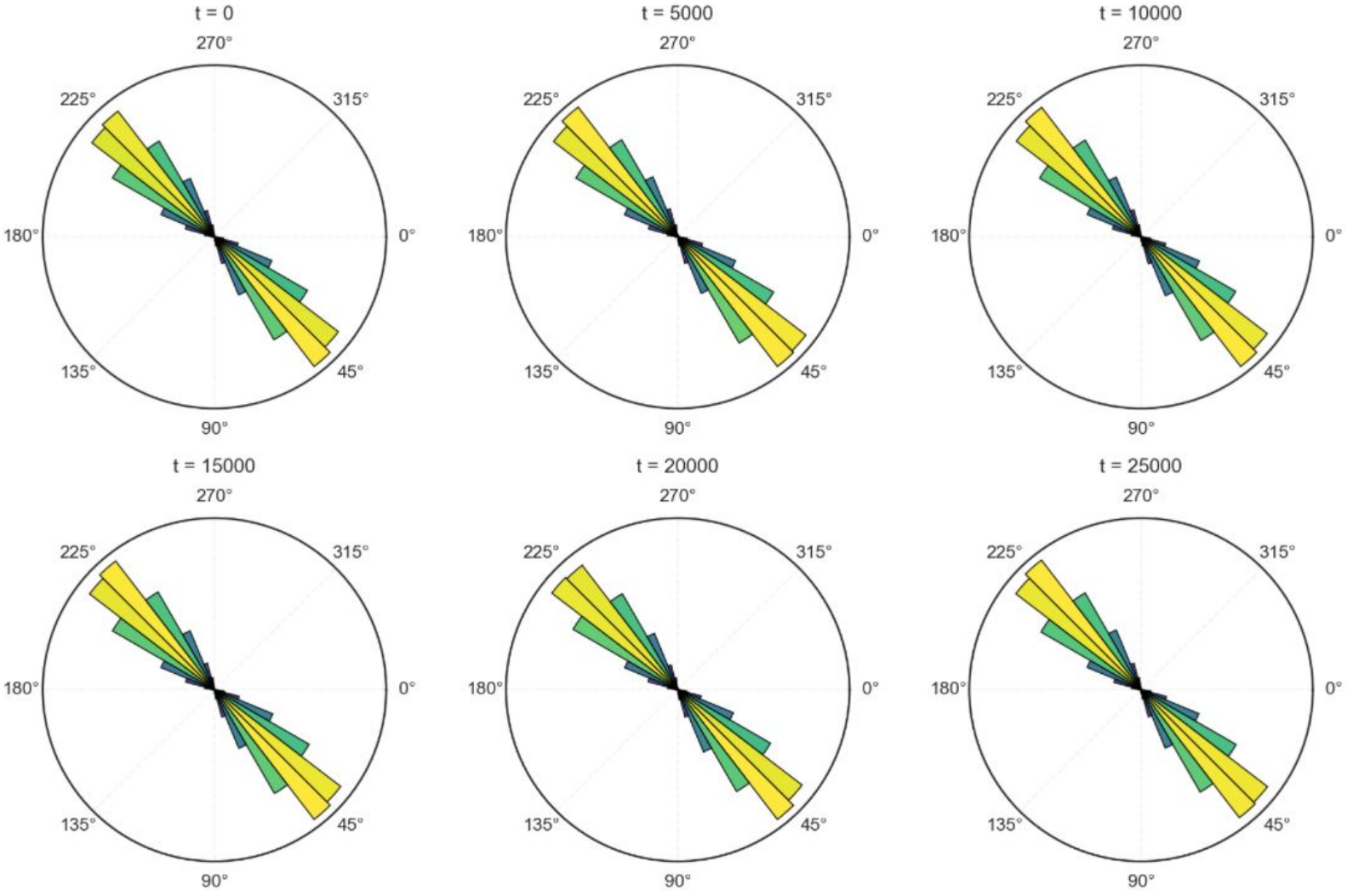
   (i) Oversmoothing: Can oversmoothing always be prevented in this model ?

   (ii) Loss of information: Do the features have any useful information after large t ?

More simulations

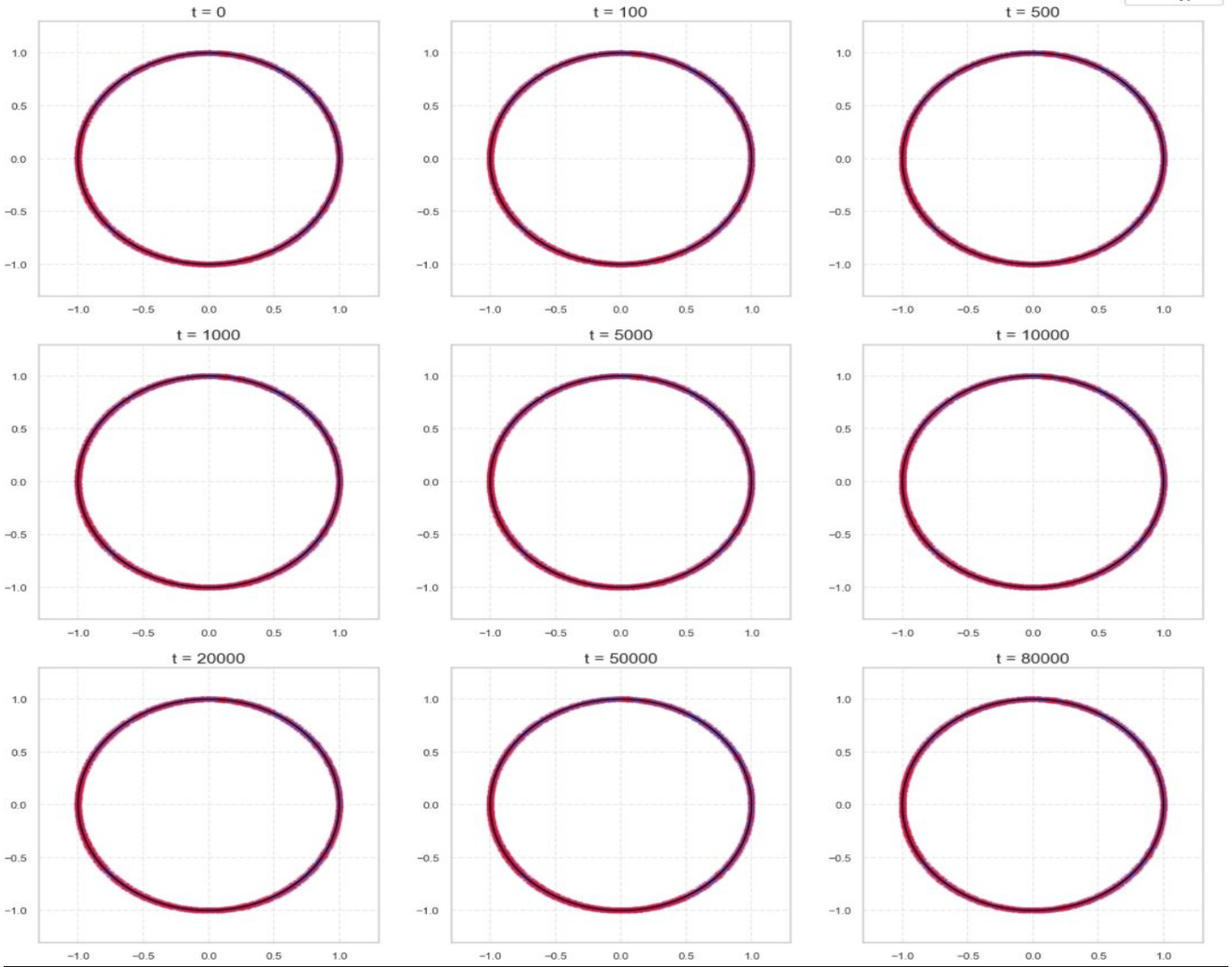
Symmetric Initial Features

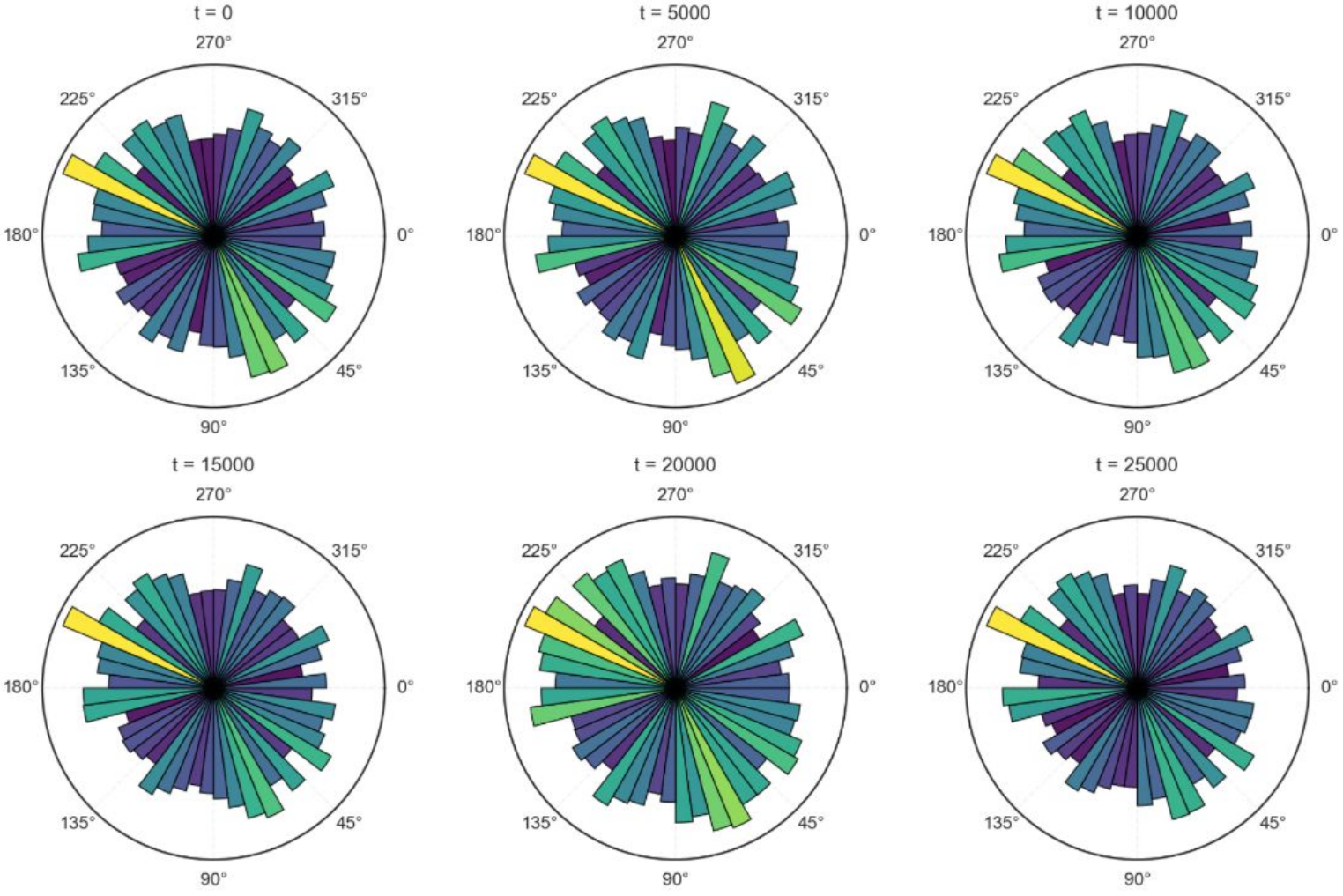



t = 0
t = 5000
t = 10000
t = 15000
t = 20000
t = 25000
270°
225°
315°
180°
0°
135°
45°
90°

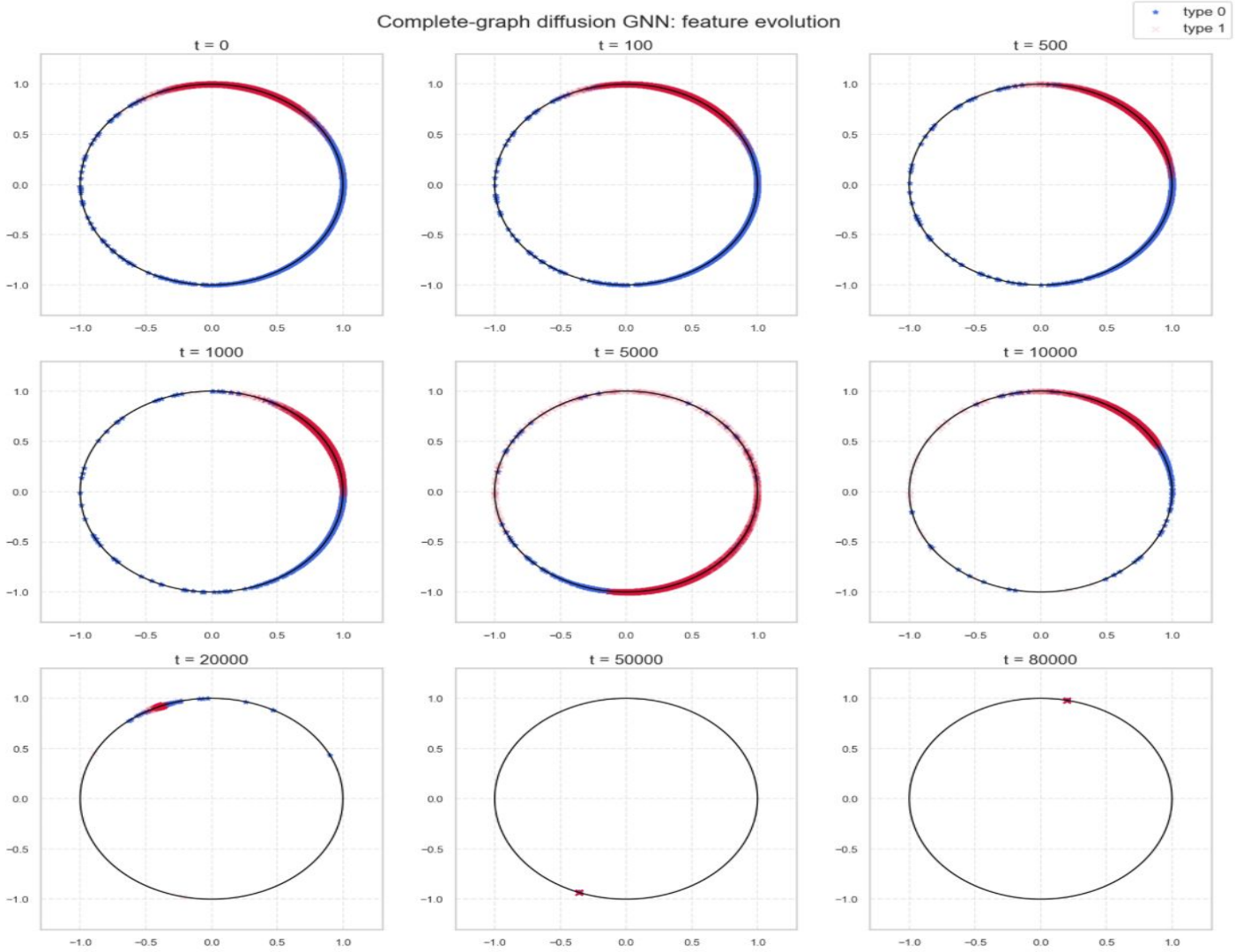
Symmetric but not well separated.

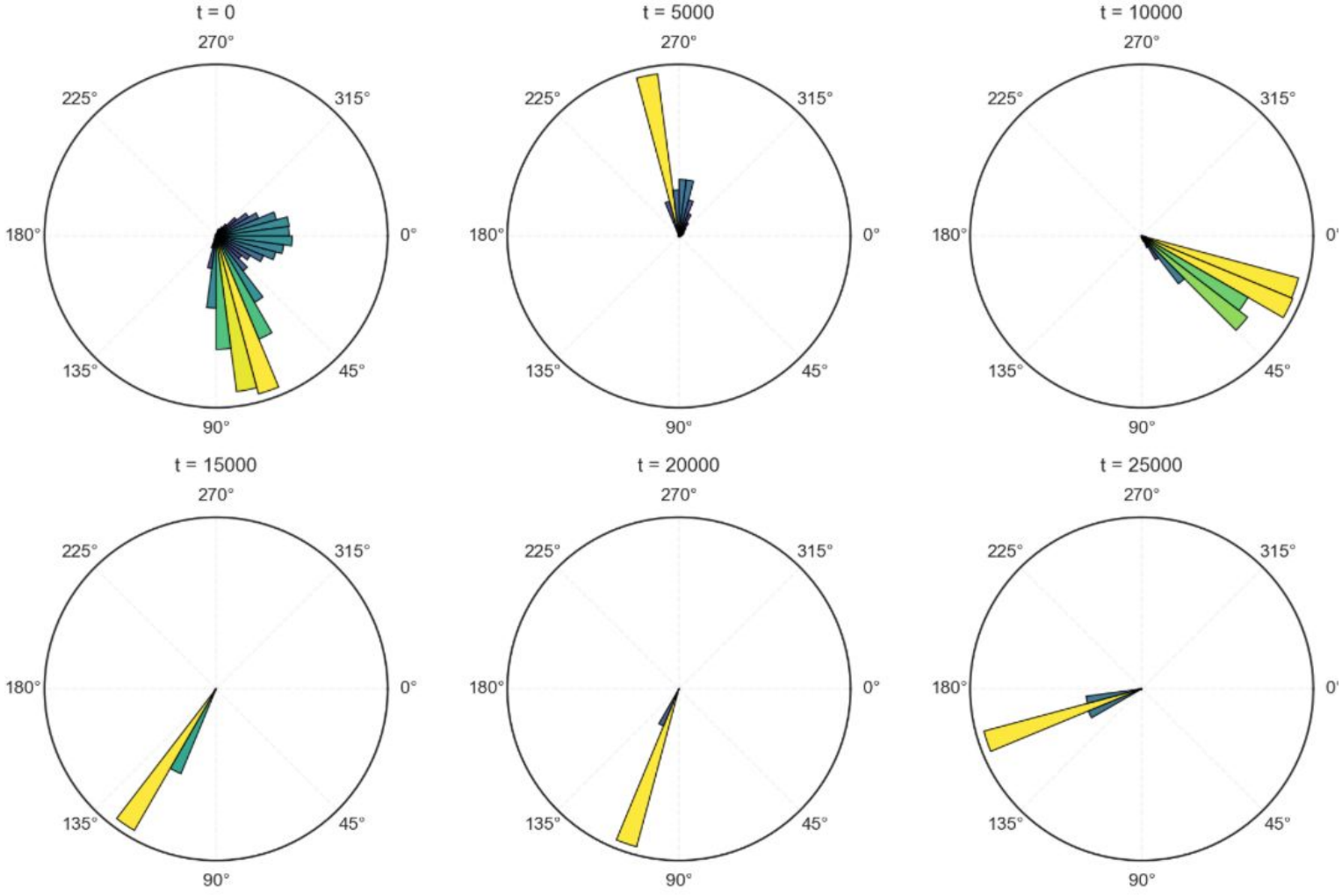
This GNN doesn’t help add any graph information.




t = 0
t = 5000
t = 10000
t = 15000
t = 20000
t = 25000
270°
225°
315°
180°
0°
135°
45°
90°

Asymmetric initial features




t = 0
t = 5000
t = 10000
t = 15000
t = 20000
t = 25000
270°
225°
315°
180°
0°
135°
45°
90°

Questions

1. So this model prevents all features collapsing to a deterministic point.
2. **Question:**

   (i) Oversmoothing: Can oversmoothing be prevented always in this model ?

   Not necessarily. When can it be avoided ?

   (ii) Loss of information: Do the features have any useful information after large t ?
3. To answer this question, we need to understand in detail **the limit of the process**

$$\lim_{t\to\infty} X_v(t) = X_v(\infty)$$

Continuous time approximation

In the limit $\alpha \to 0$, we can approximate the process by the following continuous time process

$$dX_v(t) = \sigma \Pi_{X_v(t)} \frac{1}{\deg(v)} \left[\sum_{u \sim v} X_u(t)^T \bigotimes I_2 \right] dB(t)$$

In the above,

1. $\Pi_x = (I_2 - xx^T)$ is the project to the tangent on the circle.
2. $\bigotimes$ is the Kronecker product
3. B(t) is standard Brownian motion in $\mathbb{R}^4$.

No deterministic fixed point

We have

$$dX_v(t) = \sigma \Pi_{X_v(t)} \frac{1}{\deg(v)} \left[\sum_{u \sim v} X_u(t)^T \bigotimes I_2 \right] dB(t)$$

1. Oversmoothing = All $X_v(t)$ converge to a point x almost surely.
2. Then x is a stationary solution.
3. If $X_v(t) = x$ is a stationary point, then ,

$$dX_v(t) = 0 = \sigma(I - xx^T) \left[x^T \otimes I_2 \right] dB(t)$$

4. However, this needs $\left[x^T \otimes I_2 \right] dB(t)$ to be a multiple of x, leading to a contradiction.

Complete graph setting (Dense graphs)

1. In a complete graph on n vertices, every vertex is connected to all other n-1 vertices.

$$dX_v(t) = \sigma \Pi_{X_v(t)} \frac{1}{n-1} \left[\sum_{u \neq v} X_u(t)^T \bigotimes I_2 \right] dB(t)$$

2. In particular, when n is large, we have

$$dX_v(t) \approx \sigma \Pi_{X_v(t)} \left[m_n(t)^T \bigotimes I_2 \right] dB(t)$$

   Where $m_n(t) = \frac{1}{n-1} \sum_v X_v(t)$.

3. For large n, all particles evolve similarly. Therefore, we study the behavior of one particle.

Complete graph setting

1. For large n, we need to study the following evolution

$$dX(t) = \sigma \Pi_{X(t)} \left[m(t)^T \bigotimes I_2 \right] dB(t)$$

   Where $m(t) = \mathbb{E}(X(t))$.

2. Let θ(t) be the angle made by X(t) with respect to the X-axis. Then, we have

$$d\theta(t) = \sigma \left\| m(t) \right\| d\beta(t).$$

   Where β is a standard Brownian motion.

3. This is a Markov process and as t grows larger, the process converges to its stationary distribution.

Stationary solutions

The distribution of θ has a density ϱ and satisfies the following PDE

$$\partial_t \rho(\theta, t) = \frac{\sigma^2}{2} \left| m(t) \right|^2 \partial_{\theta\theta} \rho(\theta, t), \qquad m(t) = \int_0^{2\pi} e^{i\theta} \, \rho(\theta, t) \, d\theta.$$

A stationary solution $\varrho(\theta,t) = \varrho(\theta)$ is independent of "t". Therefore, we must have

$$\left| m \right|^2 \partial_{\theta\theta} \rho(\theta) = 0 \quad \text{where } m = \int_0^{2\pi} e^{i\theta} \rho(\theta) d\theta$$

Observation 1: Any symmetric distribution is a stationary distribution as m(t) = 0.

For ex, uniform distribution.

Non-symmetric stationary solutions

Then $|m| > 0$ the diffusion coefficient $\frac{\sigma^2}{2}|m|^2$ is strictly positive, so we must have

$$\partial_{\theta\theta}\rho(\theta) = 0,$$

which implies

$$\rho(\theta) = A\theta + B$$

for some constants $A, B \in \mathbb{R}$. Since ρ is 2π-periodic, A must be zero, so ρ is constant:

$$\rho(\theta) \equiv B.$$

Normalization $\int_0^{2\pi} \rho \, d\theta = 1$ gives $B = \frac{1}{2\pi}$. But then

$$m = \int_0^{2\pi} e^{i\theta} \frac{1}{2\pi} \, d\theta = 0,$$

which contradicts $|m| > 0$. Hence no stationary solution can satisfy $|m| > 0$.

Stability of solutions

Since all symmetric distributions are stationary, we next focus on stability.

The system converges to most stable stationary distribution.

To do that, recall $\partial_t \rho(\theta,t) = \frac{\sigma^2}{2} |m(t)|^2 \partial_{\theta\theta}\rho(\theta,t), \qquad m(t) = \int_0^{2\pi} e^{i\theta}\,\rho(\theta,t)\,d\theta.$

Let $m(t) = a_1(t) + ib_1(t), \qquad r(t)^2 := |m(t)|^2 = a_1(t)^2 + b_1(t)^2.$

Stability

First, write $\rho(\theta,t) = \frac{1}{2\pi}\left(1 + 2\sum_{k=1}^{\infty}\left(a_k(t)\cos(k\theta) + b_k(t)\sin(k\theta)\right)\right)$

We then have $\frac{d}{dt}a_k(t) = -\frac{\sigma^2}{2}k^2 r(t)^2 a_k(t), \qquad \frac{d}{dt}b_k(t) = -\frac{\sigma^2}{2}k^2 r(t)^2 b_k(t)$

Therefore, $a_k(t) = a_k(0)\,(1+\sigma^2 r_0^2 t)^{-k^2/2}, \qquad b_k(t) = b_k(0)\,(1+\sigma^2 r_0^2 t)^{-k^2/2}.$

Therefore, we have $\rho(\theta,t) \longrightarrow \frac{1}{2\pi} \qquad \text{as } t \to \infty.$

Only stable stationary solution is uniform distribution.

Takeaway: complete graph stationary profile

1. Stationary solutions = Symmetric distributions
2. Only stable stationary distribution is uniform distribution

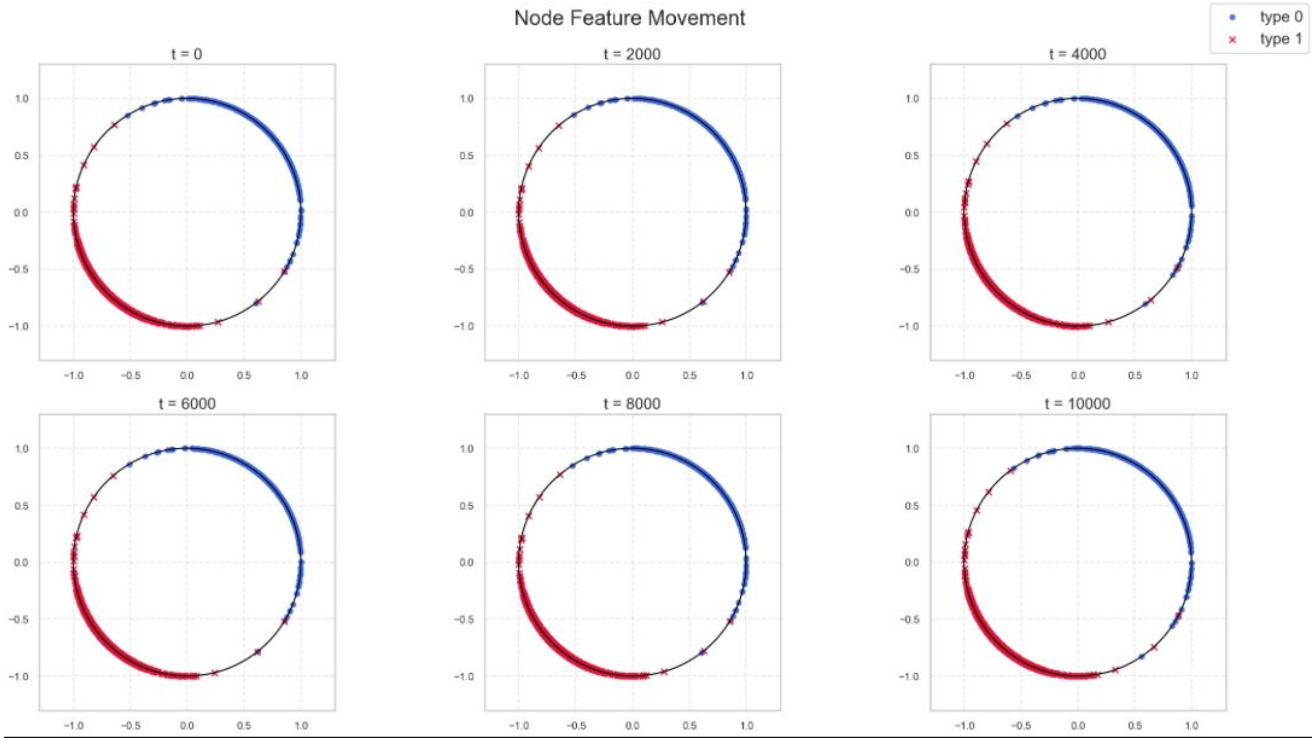
Intuitively,

1. If the features initially are symmetric, then they will be symmetric and separated after passing through GNNs.

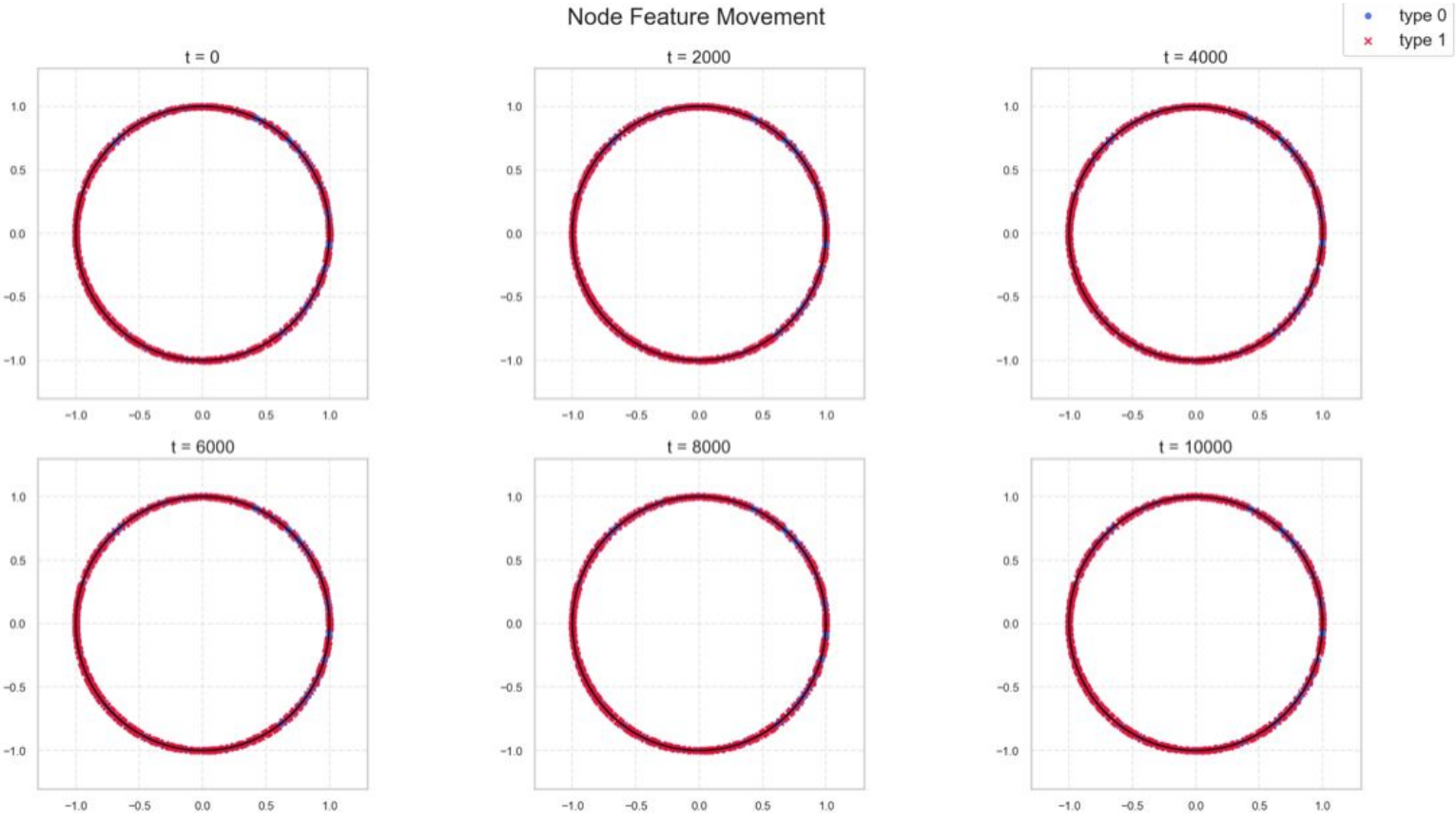
2. However, if the features deviate slightly from above, then feature information is lost completely.

Dense SBMs

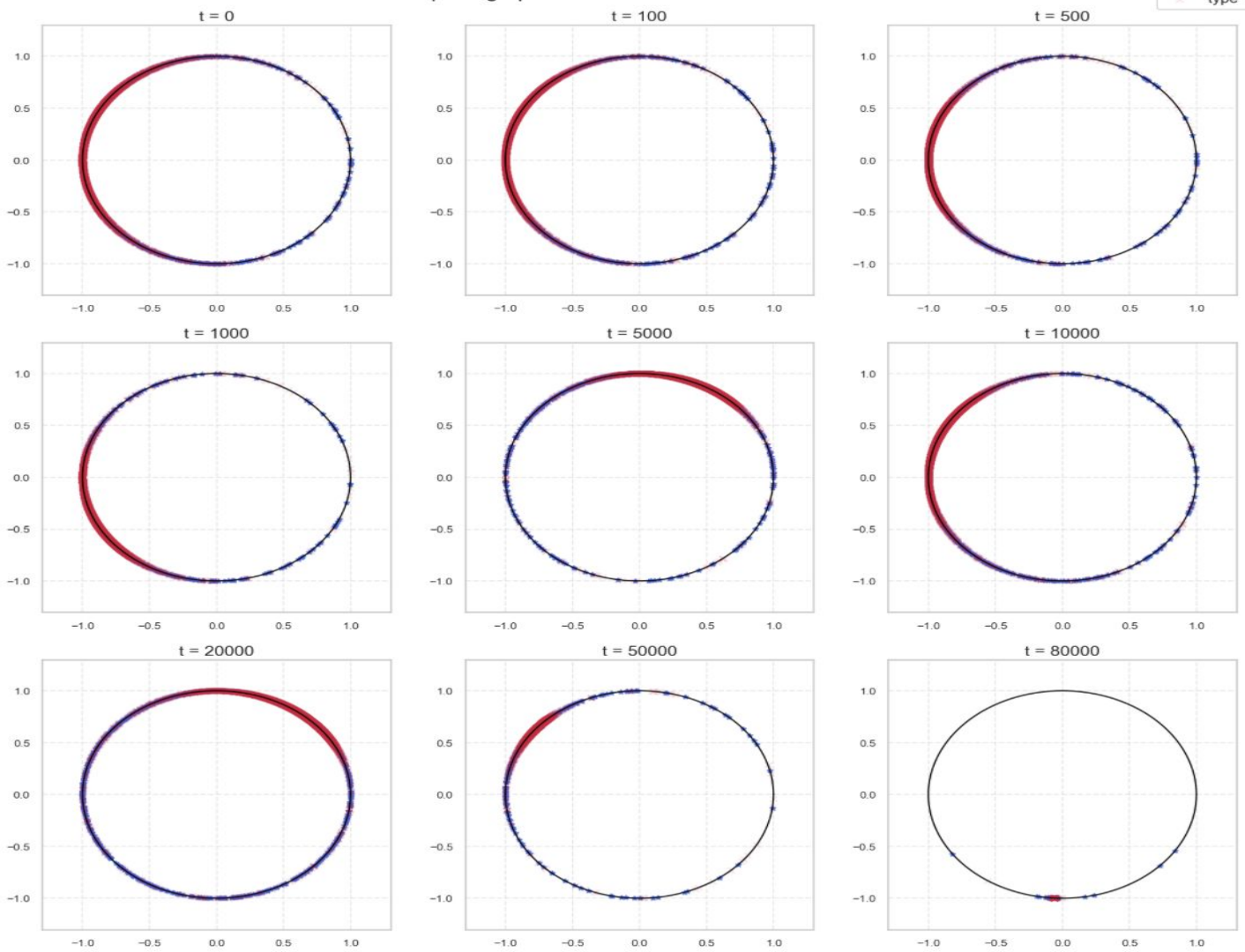
1. Well separated, symmetric



2. Symmetric and not well separated



Asymmetric features



Conclusion: Dense graphs

1. Residual GNNs with normalisation help when the features initially are well symmetric and well separated.
2. If symmetric and not well separated, oversmoothing is prevented, but feature information is lost (they become iid uniform random variables)
3. If asymmetric, then all features collapse to a “single” uniform random variable.