ACMP: ALLEN-CAHN MESSAGE PASSING WITH ATTRACTIVE AND REPULSIVE FORCES FOR GRAPH NEURAL NETWORKS

Yuelin Wang*
Shanghai Jiao Tong University
sjtu_wyl@sjtu.edu.cn

Kai Yi*
University of New South Wales
kai.yi@unsw.edu.au

Xinliang Liu*
King Abdullah University of Science and Technology
xinliang.liu@kaust.edu.sa

Yu Guang Wang†
Shanghai Jiao Tong University
yuguang.wang@sjtu.edu.cn

Shi Jin†
Shanghai Jiao Tong University
shijin-m@sjtu.edu.cn

ABSTRACT

Neural message passing is a basic feature extraction unit for graph-structured data considering neighboring node features in network propagation from one layer to the next. We model such process by an interacting particle system with attractive and repulsive forces and the Allen-Cahn force arising in the modeling of phase transition. The dynamics of the system is a reaction-diffusion process which can separate particles without blowing up. This induces an Allen-Cahn message passing (ACMP) for graph neural networks where the numerical iteration for the particle system solution constitutes the message passing propagation. ACMP which has a simple implementation with a neural ODE solver can propel the network depth up to one hundred of layers with theoretically proven strictly positive lower bound of the Dirichlet energy. It thus provides a deep model of GNNs circumventing the common GNN problem of oversmoothing. GNNs with ACMP achieve state of the art performance for real-world node classification tasks on both homophilic and heterophilic datasets. Codes are available at https://github.com/ykiiiiii/ACMP.

1 INTRODUCTION

Graph neural networks (GNNs) have received a great attention in the past five years due to its powerful expressiveness for learning graph structured data, with broad applications from recommendation systems to drug and protein designs (Atz et al., 2021; Baek et al., 2021; Bronstein et al., 2021; 2017; Gainza et al., 2020; Wu et al., 2020). Neural message passing (Gilmer et al., 2017) serves as a fundamental feature extraction unit for graph-structured data that aggregates the features of neighbors in network propagation. We develop a GNN message passing, called the *Allen-Cahn message passing (ACMP)*, using interacting particle dynamics, where nodes are particles and edges representing the interactions of particles. The system is driven by both attractive and repulsive forces, plus the Allen-Cahn double-well potential from phase transition modeling. This model is motivated by the

*equal contribution
†corresponding author

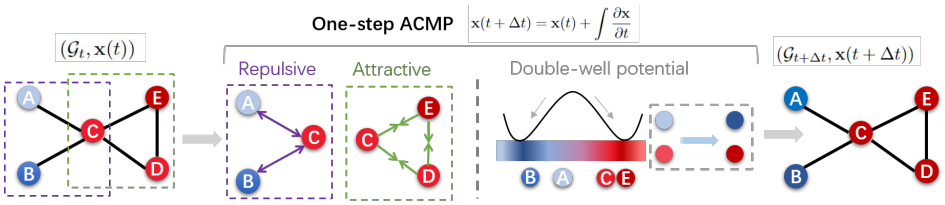

Figure 1: An illustration for one-step ACMP. Graph $\mathcal{G}_t$ with features $\mathbf{x}(t)$ in the purple and green blocks have different treatment of attraction or repulsion. The same color indicates similar node features. The node $\mathbf{x}(t)$ is updated by one step to $\mathbf{x}(t+\Delta t)$ via ODE solver. Nodes in the green block tend to attract each other and in the other block, nodes in different colors repel each other, and thus both colors are strengthened during propagation. It gives rise to forming bi-cluster flocking. The double-well potential turns features darker under gradient flow to circumvent blowup of energy.

behavior of the particle system of collective behaviors common in nature and human society, for example, insects forming swarms to work; birds forming flocks to immigrate; humans forming parties to express public opinions. Various mathematical models have been proposed to model these behaviors (Albi et al., 2019; Motsch & Tadmor, 2011; Castellano et al., 2009; Proskurnikov & Tempo, 2017; Degond & Motsch, 2008). There are two major components in this model. First, while the attractive force forces all particles into one cluster, the repulsive forces allow particles to separate into two different clusters, which is essential to avoid oversmoothing. However, repulsive forces could make the Dirichlet energy diverge. We augment the model with the Allen-Cahn (Allen & Cahn, 1979) term (or Rayleigh friction (Rayleigh, 1894)), which is crucial in preventing the Dirichlet energy in the evolution from becoming unbounded, allowing us to prove mathematically that the lower bound of the Dirichlet energy is strictly bigger than zero, hence avoiding oversmoothing. Specifically, we will prove that under suitable conditions on the parameters, the dynamics of the ACMP particle system will time-asymptotically form 2^d different clusters and the Dirichlet energy has a strictly positive lower bound.

The structure of ACMP can handle two problems in GNNs: oversmoothing and heterophily. Oversmoothing (Nt & Maehara, 2019; Oono & Suzuki, 2019; Konstantin Rusch et al., 2022) means that all node features become undistinguishable, and equivalently, in the formulation of particle systems, features form only one consensus. Heterophily problems means GNNs perform worse in heterophilic graphs (Lim et al., 2021; Yan et al., 2021). It is due to the neighboring nodes of different classes are mistaken for the same class in GNNs like GCN and GAT. However, the presence of repulsion in ACMP makes particles separate into two different clusters, hence provides a simple and neat solution for prediction tasks on both two problems.

Overall, the benefit of the Allen-Cahn message passing with repulsion is manifold. 1) It circumvents oversmoothing issue, namely the Dirichlet energy is bounded from below. 2) The network is stable in the sense that features and Dirichlet energy are bounded from above. 3) Feature smoothness (energy decreasing) and the balance between nodes features and edge features can be adjusted by network parameters that control the attraction, repulsion and phase transition. The model can then reach an acceptable trade-off on self-features and neighbor effect, as shown in Figure 1. Our model can thus handle node classification tasks for both homophilic and heterophilic datasets by using only one-hop neighbour information. 4) The proposed model can be implemented by neural ODE solvers for the system with attractive and repulsive forces.

In theory, we prove that Dirichlet energy of GNNs with ACMP has a lower bound above zero (limiting oversmoothing), as well as an upper bound (circumventing blow-up) under specific conditions. This agrees with the experimental results (Section 6). We also prove that ACMP is a process for the features to generate clusters thanks to the double-well potential, which provides an interpretable theory for node classification.

2 Background

Message Passing in GNNs Graph neural networks are a kind of deep neural networks which take graph data as input. Neural Message Passing (MP) (Gilmer et al., 2017; Battaglia et al., 2018) is a

most widely used propagator for node feature update in GNNs, which takes the following form: for the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is with sets of nodes $\mathcal{V}$ and edges $\mathcal{E}$, with $\mathbf{x}_i^{(k-1)} \in \mathbb{R}^d$ denoting features of node i in layer $(k-1)$ and $a_{j,i} \in \mathbb{R}^D$ edge features from node j to node i,

$$\mathbf{x}_i^{(k)} = \gamma^{(k)} \left(\mathbf{x}_i^{(k-1)}, \square_{j \in \mathcal{N}_i} \, \phi^{(k)} \left(\mathbf{x}_i^{(k-1)}, \mathbf{x}_j^{(k-1)}, a_{j,i} \right) \right),$$

where $\square$ denotes a differentiable, (node) permutation invariant function, e.g., sum, mean or max, and γ and ϕ denote differentiable functions such as MLPs (MultiLayer Perceptrons), and $\mathcal{N}_i$ is the set of one-hop neighbors of node i. The message passing updates the feature of each node by aggregating the self-feature with neighbors' features. Many GNN feature extraction modules such as GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018) and GIN (Xu et al., 2018) can be written as message passing. For example, the MP of GCNs reads, with learnable parameter matrix $\mathbf{\Theta}$, $\mathbf{x}_i' = \mathbf{\Theta}^\top \sum_{j \in \mathcal{N}_i \cup \{i\}} \frac{a_{j,i}}{\sqrt{\hat{d}_j \hat{d}_i}} \mathbf{x}_j$, where $\hat{d}_i = 1 + \sum_{j \in \mathcal{N}(i)} a_{j,i}$ and $\hat{D} = \mathrm{diag}(\hat{d}_1, \ldots, \hat{d}_N)$ is the degree matrix for $A + I$. Graph attention network (GAT) uses attention coefficients $\alpha_{i,j}$ as similarity information between nodes in the MP update $\mathbf{x}_i' = \alpha_{i,i} \mathbf{\Theta} \mathbf{x}_i + \sum_{j \in \mathcal{N}_i} \alpha_{i,j} \mathbf{\Theta} \mathbf{x}_j$, with

$$\alpha_{i,j} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{\Theta}\mathbf{x}_i \,\|\, \mathbf{\Theta}\mathbf{x}_j]\right)\right)}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top [\mathbf{\Theta}\mathbf{x}_i \,\|\, \mathbf{\Theta}\mathbf{x}_k]\right)\right)}. \tag{1}$$

The MP framework was also developed as PDE solvers in Brandstetter et al. (2022b) by embedding differential equations as a parameter into message passing like Brandstetter et al. (2022a). This paper regards particle system evolution (ODE) as message passing propagation, and the appropriate design of the particle system offers desired properties for the resulting GNN.

Graph neural diffusion Neural diffusion equations on graphs (GRAND) are proposed by Chamberlain et al. (2021), which provides a unified mathematical framework for some message passings:

$$\frac{\partial}{\partial t} \mathbf{x}(t) = \mathrm{div}[\mathbf{G}(\mathbf{x}(t), t) \nabla \mathbf{x}(t)], \tag{2}$$

where $\mathbf{G} = \mathrm{diag}(a(x_i(t), x_j(t), t))$ where a is a function reflecting similarity between nodes i and j, and x_i is the scale-valued feature for node i, and $\mathbf{x} = \oplus x_i$.

3 Motivations

3.1 Attractive and repulsive forces

The equation (2) itself can be interpreted in a formulation different from diffusion. In this paper, we study the neural equations of *interacting particle system*, which has a similar structure to (2). We rewrite (2) into a component-wise version and obtain a particle system

$$\frac{\partial}{\partial t} x_i(t) = \sum_{j \in \mathcal{N}_i} a(x_i, x_j)(x_j - x_i). \tag{3}$$

In the formulation of particle systems, one can easily discover the evolution trend of the features. If $a(x_i, x_j) > 0$, the direction of x_i's velocity is towards x_j, which means that x_i is attracted by x_j. In the contrast, if $a(x_i, x_j) < 0$, x_i has a trend to move away from x_j. Hence, $a(x_i, x_j)$ serves as the attractiveness or repulsiveness of the force between x_i and x_j. In the diffusion model above, all $a(x_i, x_j)$'s are positive, therefore all the node features in one connected component attract each other. If the weight matrix $(a(x_i, x_j))_{N \times N}$ is right-stochastic, one can prove that the convex hull of the features will not dilate in time (see Motsch & Tadmor (2014); Chamberlain et al. (2021)). Such feature aggregation means that message propagates along the edges of the graph and some potential *consensus* forms in the process.

However, the message propagation does not limit to *consensus* (corresponding to diffusion). Information interaction can derive polarization of final judgement when *negative* message matters in some problems rather than *positive* message. For instance, in a node classification task on a bipartite, the neighbour message is negative since connected nodes belong to different classes. In the

formulation of particle systems, the mechanism of positive and negative messages can be modelled by adding bias $\beta_{i,j}$ into (3)

$$\frac{\partial}{\partial t}x_i(t) = \sum_{j\in\mathcal{N}_i}(a(x_i,x_j)-\beta_{i,j})(x_j-x_i). \tag{4}$$

The coefficient term $a(x_i,x_j)-\beta_{i,j}$ corresponds to the interactive force. By adjusting $\beta_{i,j}$, both in the system attractive and repulsive forces co-exist. If $a(x_i,x_j)-\beta_{i,j}>0$, x_i is attracted by x_j. While if $a(x_i,x_j)-\beta_{i,j}<0$, x_i is repelled by x_j. If the coefficient equates zero, there is no interaction between x_i and x_j. Then, the dynamics is enabled to adapt both positive and negative message passing. In this way, the neural message passing can handle either homophilic or heterophilic datasets (see Section 6 for detailed discussion).

3.2 Pseudo-Ginzburg-Landau energy

However, adding the repulsion term may cause the particles being separated away infinitely, thus the Dirichlet energy becomes unbounded. To avoid this problem, we add a damping term $\delta x_i(1-x_i^2)$, which we call an *Allen-Cahn term*. The coefficient $\alpha>0$ is multiplied just for technical convenience.

$$\frac{\partial}{\partial t}x_i(t) = \alpha\sum_{j\in\mathcal{N}_i}(a(x_i,x_j)-\beta_{i,j})(x_j-x_i)+\delta x_i(1-x_i^2). \tag{5}$$

Gradient Flow The variational principle governing many PDE models states that the equilibrium state is actually the minimizer of one specific energy. We first introduce the Dirichlet energy and show that (3) can be characterized by looking into the corresponding Euler-Lagrange equation of the Dirichlet energy. Let adjacent matrix $\mathbf{A}$ represent the undirected connectivity between nodes x_i and x_j, with $a_{i,j}=1$ for $(i,j)\in\mathcal{E}$ and $a_{i,j}=0$ for $(i,j)\notin\mathcal{E}$. The *Dirichlet energy* $\mathbf{E}$ in terms of $\mathcal{G}=(\mathcal{V},\mathcal{E})$ and node features $\mathbf{x}\in\mathbb{R}^{N\times d}$ takes the form

$$\mathbf{E}(\mathbf{x}) = \frac{1}{N}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i}a_{i,j}\|\mathbf{x}_i-\mathbf{x}_j\|^2. \tag{6}$$

By calculus of variation, we can formulate the corresponding particle equation

$$\frac{\partial\mathbf{x}}{\partial t} = -\nabla_{\mathbf{x}}\mathbf{E}, \quad \frac{\partial x_i}{\partial t} = -\frac{\partial\mathbf{E}}{\partial x_i} = \frac{2}{N}\sum_{j\in\mathcal{N}_i}a_{i,j}(x_j-x_i). \tag{7}$$

On the RHS of (7), the summation takes over the one-hop neighbors $\mathcal{N}_i$ of node i, which aggregates the impact from the neighboring nodes. Equation (7) is (5) when one takes adjacent matrix $\mathbf{A}$ as the weight matrix $(a(x_i,x_j))_{N\times N}$.

Particle equation with the double-well potential To avoid blowing-up of the solution, one can design an external potential to control the solutions so they are bounded. Here, we define the *pseudo-Ginzburg-Landau energy* on graph $\mathcal{G}$ denoted by $\Phi: L^2(\mathcal{V})\to\mathbb{R}$, as a combination of the *interacting energy* and *double-well potential* $W:\mathbb{R}\to\mathbb{R}_+$, with

$$W(x) = (\delta/4)(1-\|x\|^2)^2, \qquad \Phi(\mathbf{x}) = \frac{1}{2}\alpha\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}_i}(a_{i,j}-\beta_{i,j})\|x_i-x_j\|^2+\sum_{i\in\mathcal{V}}W(x_i), \tag{8}$$

where parameters $\alpha,\delta>0$ are used to balance the two types of energy. From now on, we denote $a(x_i,x_j)$ by $a_{i,j}$ for simplicity. The pseudo-Ginzburg-Landau energy is not a true energy because the matrix $(a_{i,j}-\beta_{i,j})_{N\times N}$ can be non-positive definite. If $\beta_{i,j}$'s all equate zero, it then becomes the Ginzburg-Landau energy defined in Bertozzi & Flenner (2012); Luo & Bertozzi (2017). Using this combined energy, we can obtain the Allen-Cahn equation with repulsion on graph as $\frac{\partial\mathbf{x}}{\partial t}=-\nabla_{\mathbf{x}}\Phi$, which is equivalent to (5).

4 Allen-Cahn Message Passing

We propose the *Allen-Cahn Message Passing* (ACMP) neural network based on equation (5), where the message is updated by the evolution of the equation via a neural ODE solver. To our best knowledge, this is the *first time* to introduce a type of message passing to amplify the difference between connected nodes by repulsive force.

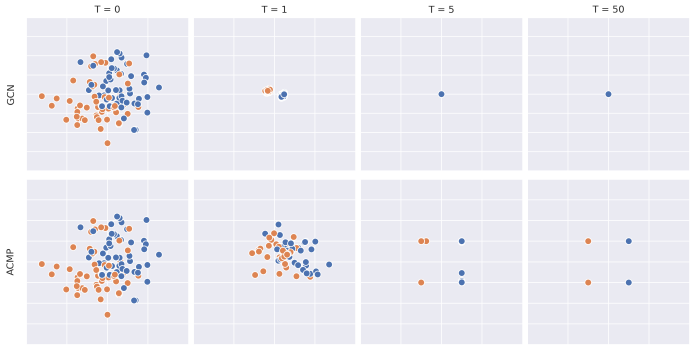

Figure 2: We compare the evolution of node features in GCN and ACMP. The initial position is represented by the 2-dimensional position of the nodes, which is shown in the first column. The GCN aggregates all node features by taking the weighted average of its neighbors' features. With the propagated steps increasing, all the nodes' features shrink to a point, which gives rise to oversmoothing. When it comes to ACMP, nodes' features are grouped by four attractors, which helps to circumvent oversmoothing. More details can be seen in Section 6.

Network Architecture Suppose d-dimensional node-wise features represented by a matrix $\mathbf{x}^{\text{in}}$ where row i represents feature of node i. Our scheme first embeds the node feature $\mathbf{x}(0) = \text{MLP}(\mathbf{x}^{\text{in}})$ by a simple multi-layer perceptron (MLP), which is treated as an input for ACMP propagation $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$, by $\mathbf{x}(0) \mapsto \mathbf{x}(T)$, where $\mathbf{x}(T) = \mathbf{x}(0) + \int_0^T \frac{\partial \mathbf{x}(t)}{\partial t} dt, \quad \mathbf{x}(0) = \text{MLP}\left(\mathbf{x}^{\text{in}}\right)$, where $\frac{\partial \mathbf{X}(t)}{\partial t}$ is estimated by ACMP defined on $\mathcal{G}$ based on (5). The node features $\mathbf{x}(T)$ at the ending time are fed into an MLP based classifier. Then, we define the Allen-Cahn message passing by

$$\frac{\partial}{\partial t}\mathbf{x}_i(t) = \boldsymbol{\alpha} \odot \sum_{j \in \mathcal{N}_i} (a(\mathbf{x}_i(t), \mathbf{x}_j(t)) - \beta)(\mathbf{x}_j(t) - \mathbf{x}_i(t)) + \boldsymbol{\delta} \odot \mathbf{x}_i(t) \odot (1 - \mathbf{x}_i(t) \odot \mathbf{x}_i(t)). \quad (9)$$

Here $\boldsymbol{\alpha}, \boldsymbol{\delta} \in \mathbb{R}^d$ are learnable vectors of the same length as the node feature $\mathbf{x}_i$. While we can use a more general case when each edge (i, j) uses different trainable $\beta_{i,j}$, we have simplied to single hyper-parameter $\beta \in \mathbb{R}^+ \cup \{0\}$, which makes the network and optimization easier. The β in our model is a crucial parameter, which can be adjusted such that the attractive and repulsive forces both present to enrich the message passing effect. If one chooses $\boldsymbol{\delta} = 0$, $\beta = 0$, our model is reduced to the graph neural diffusion network (GRAND) in Chamberlain et al. (2021). In experiments, we would make significant use of nontrivial $\boldsymbol{\delta}$ and β.

The operations of all terms are channel-wise, involving d channels, except $a(\mathbf{x}_i(t), \mathbf{x}_j(t))$, and $\odot$ represents channel-wise multiplication for d feature channels. Figure 1 illustrates the one-step ACMP mechanism (9): Nodes with close colors attracts each other otherwise repel. Nodes in the same block tend to attract each other and both colors are strengthened during message passing propagation. The double-well potential prevents the features and Dirichlet energy from blowup. In this process, node feature $\mathbf{x}(t)$ is updated to $\mathbf{x}(t + \Delta t)$ for a time increment Δt. Ultimately, a bi-cluster flock is formed for node classification.

In the propagation of ACMP in (9), we need to specify how the neighbors are interacted, that is how the $a(\mathbf{x}_i(t), \mathbf{x}_j(t))$ is evolved with time. There are many kinds of methods to update the edge weights. The two typical types are GCN (Kipf & Welling, 2017) and GAT (Veličković et al., 2018).

ACMP-GCN: this model uses deterministic $a(\mathbf{x}_i(t), \mathbf{x}_j(t))$, which is given by the adacency matrix $A = (a_{i,j})$ of the original input graph $\mathcal{G}$ and does not change with time. That is, the coefficients in GCNs $a_{i,j}^{\text{GCN}} := a_{i,j}/\sqrt{\hat{d}_i \hat{d}_j}$. The message passing of (9) is reduced to

$$\frac{\partial}{\partial t}\mathbf{x}_i(t) = \boldsymbol{\alpha} \odot \sum_{j \in \mathcal{N}_i} (a_{i,j}^{\text{GCN}} - \beta)(\mathbf{x}_j(t) - \mathbf{x}_i(t)) + \boldsymbol{\delta} \odot \mathbf{x}_i(t) \odot (1 - \mathbf{x}_i(t) \odot \mathbf{x}_i(t)). \quad (10)$$

ACMP-GAT: we can replace $a_{i,j}^{\mathrm{GCN}}$ in (10) by the attention coefficients (1) of GAT, which with extra trainable parameters measures the similarity between two nodes by taking account of both node and structure features. The system then drives edges to update in each iteration of message passing.

Neural ODE Solver Our method uses an ODE solver to numerically solving the equation ((9) and (10)). To obtain the node features $\mathbf{x}(T)$, we need a stable numerical integrator for solving the ODE efficiently and backpropagation of gradients. Since our model is stable in terms of evolution time, most explicit and implicit numerical methods such as explicit Euler, Runge-Kutta 4th-order, midpoint, Dormand-Prince5 (Chen et al., 2018; Lu et al., 2018; Norcliffe et al., 2020; Chamberlain et al., 2021) work well as long as the step size τ is small enough. In experiments, we implement ACMP using Dormand-Prince5 method which provides a fast and stable numerical solver. The *network depth* of ACMP-GNN is equal to the numerical iteration number n_t set in the solver.

Computational Complexity The computational complexity of the ACMP is $\mathcal{O}(NEdn_t)$, where n_t, N, E and d are number of time steps in time interval $[0, T]$, number of nodes, number of edges and number of feature dimension, respectively. Since our model only considers nearest (one-hop) neighbors, E is significantly smaller than that of graph rewiring (Gasteiger et al., 2019; Alon & Yahav, 2021) and multi-hop (Zhu et al., 2020) methods.

Channel Mixer Channel mixing can be spontaneously introduced from the perspective of diffusion coefficients though our model is previously written in the channel-wise form. Whether channel mixing happens depends on the specific GNN driver we choose for ACMP. When the coefficients $\mathbf{a}(\mathbf{x}_i(t), \mathbf{x}_j(t))$ in (9) that do not update with time are a scalar or vector, like in ACMP-GCN, the operations of the message passing propagator are channel-wise and channel mixing is not incorporated. On the other hand, the ACMP-GAT with graph attention driver incorporates a learnable channel mixing when the coefficients are tensors. The channel mixer can be introduced by generalizing the Dirichlet energy to high dimension, for example, $\mathbf{E}(\mathbf{x}) := \frac{1}{N} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{a}_{i,j} (\mathbf{x}_i - \mathbf{x}_j)$, when $\mathbf{a}_{i,j} \in \mathbb{R}^{d \times d}$ are connectivity tensors.

5 Dirichlet Energy

The dynamics (5) can circumvent the oversmoothing issue of GNNs (Nt & Maehara, 2019; Oono & Suzuki, 2019; Konstantin Rusch et al., 2022). Oversmoothing phenomenon means that all node features converge to the same constant – consensus forms – as the network deepens, and equivalently, the Dirichlet energy will decay to zero exponentially. This idea was first introduced in Cai & Wang (2020). Konstantin Rusch et al. (2022) gives an explicit form for oversmoothing.

In our model, as we will show below, the node features in each channel tend to evolve into two clusters departing from each other under certain conditions. This implies a strictly positive lower bound of the Dirichlet energy. In addition, the system will not blow up thanks to the Allen-Cahn term. We put all the proofs and some related supplementary results in the appendix.

Proposition 1. *If $\delta > 0$, the node features $\mathbf{x}_i$ in (5) is bounded in terms of $\|\cdot\|$ and energy for all $t > 0$, i.e., $\mathbf{E}(\mathbf{x}(t)) \leq C$, and $\|\mathbf{x}\| \leq C$, where the constant C only depends on N and $\lambda_{\max}$.*

In the following propositions, we imitate the emergent behavior analysis in Fang et al. (2019) (see Appendix for details). For a graph $\mathcal{G}$ with N nodes, its vertices are said to form *bi-cluster flocking* if there exist two disjoint sets of vertex subsets $\{\mathbf{x}_i^{(1)}\}_{i=1}^{N_1}$ and $\{\mathbf{x}_i^{(2)}\}_{j=1}^{N_2}$ satisfying

$$\begin{aligned} &(i) \sup_{0 \leq t < \infty} \max_{1 \leq i,j \in N_1} |x_i^{(1)}(t) - x_j^{(1)}(t)| < \infty, \quad \sup_{0 \leq t < \infty} \max_{1 \leq i,j \in N_2} |x_i^{(2)}(t) - x_j^{(2)}(t)| < \infty; \\ &(ii) \ \exists\, C', T^{**} > 0 \textit{ such that } \min_{1 \leq i \in N_1, 1 \leq j \in N_2} \{|x_i^{(1)}(t) - x_j^{(2)}(t)|\} \geq C', \quad \forall t > T^{**}, \end{aligned} \tag{11}$$

where $x_i^{(1)}, x_i^{(2)}$ denote any component of $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}$.

We now show the long-time behaviour of model (5) following the analysis of Fang et al. (2019) for strength coupling (α, δ) that satisfies the following condition: there exists $\{\beta_{i,j}\}$ such that $\mathcal{I} :=$

$\{1, \ldots, N\}$ can be divided into two disjoint groups $\mathcal{I}_1, \mathcal{I}_2$ with N_1 and N_2 particles respectively:

$$\begin{aligned} 0 < S \leq \overline{a}_{i,j} \ \text{ with } \overline{a}_{i,j} &:= a_{i,j} - \beta_{i,j} \qquad \text{for } i,j \in \mathcal{I}_1, \\ 0 < S \leq \overline{a}_{i,j} \ \text{ with } \overline{a}_{i,j} &:= a_{i,j} - \beta_{i,j} \qquad \text{for } i,j \in \mathcal{I}_2, \\ 0 \leq \overline{a}_{i,j} \leq D \ \text{ with } \overline{a}_{i,j} &:= -(a_{i,j} - \beta_{i,j}) \qquad \text{otherwise,} \end{aligned} \tag{12}$$

where S, D are independent of time t. The S and D in (12) are the repulsive and attractive forces. We prove that if the repulsive force between the particles is stronger than the attractive force, that is, $S > D$, the system is guaranteed to have bi-cluster flocking, as shown in Proposition 2 below. For time $t \geq 0$, suppose $\mathbf{x}_c^{(1)}(t)$ and $\mathbf{x}_c^{(2)}(t)$ are the *feature centers* of the two groups of the particles $\{\mathbf{x}_i^{(1)}(t)\}_{i=1}^{N_1}$ and $\{\mathbf{x}_j^{(2)}(t)\}_{j=1}^{N_2}$ which are partitioned as above from the whole vertex set $\mathcal{V}$, given by

$$\mathbf{x}_c^{(1)}(t) := \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbf{x}_i^{(1)}(t), \quad \mathbf{x}_c^{(2)}(t) := \frac{1}{N_2} \sum_{i=1}^{N_2} \mathbf{x}_i^{(2)}(t).$$

Suppose $\mathbf{x}_c^{(s)}(t)$ has the d-dimensional feature, and let $\mathbf{x}_{c,k}^{(s)}(t)$, $k = 1, \ldots, d$, be the kth (dimension) component of the feature $\mathbf{x}_c^{(s)}(t)$, $s = 1, 2$.

Proposition 2. *The system (5) has a bi-cluster flocking if for each $k = 1, \ldots, d$, the initial $|\mathbf{x}_{c,k}^{(1)}(0) - \mathbf{x}_{c,k}^{(2)}(0)| \gg 1$, and if there exists a positive constant η such that*

$$\alpha(S - D) \min\{N_1, N_2\} \geq \delta + \eta, \tag{13}$$

where the δ is the weight factor for the double-well potential in the equation (5).

Proposition 3. *For system (5) with bi-cluster flocking, there exists a constant $C > 0$ and some time T^* such that $\forall t \geq T^*$,*

$$|\mathbf{x}_i^{(1)}(t) - \mathbf{x}_j^{(2)}(t)| \geq C > 0, \quad \forall i, j.$$

Thus, if the non-zero $a_{i,j}$ are all positive, the Dirichlet energy for ACMP is lower bounded by a positive constant.

6 Experiments

Dirichlet Energy We first illustrate the evolution of the Dirichlet energy of ACMP by an undirected synthetic random graph. The synthetic graph has 100 nodes with two classes and 2D feature which is sampled from the normal distribution with the same standard deviation $\sigma = 2$ and two means $\mu_1 = -0.5$, $\mu_2 = 0.5$. The nodes are connected randomly with probability $p = 0.9$ if they are in the same class, otherwise nodes in different classes are connected with probability $p = 0.1$. We compare the performance of GNN models with four message passing propagators: GCNs (Kipf & Welling, 2017), GAT (Veličković et al., 2018), GRAND (Chamberlain et al., 2021) and ACMP-GCN. In Figure 2, we visualize how the node features evolve from their initial state to their final steady state when 50 layers of GNN are applied. Additionally, in Figure 3, we show the Dirichlet energy of each layer's output in logarithm scales. Traditional GNNs such as GCNs and GAT suffer oversmoothing as the Dirichlet energy exponentially decays to zero in the first ten layers. GRAND relieves this problem by multiplying a small constant which can delay all nodes' features to collapse to the same value. For ACMP, the energy stabilizes at the level that relies upon the roots of the double-well potential in (9) after slightly decaying in the first two layers.

Node Classification We compare the performance of ACMP with several popular GNN model architectures on various node classification benchmarks, containing both homophilic and heterophilic datasets. Graph data is considered as *homophilic* (Pei et al., 2020) if similar nodes in the graph tend to connect together. Conversely, the graph data is said *heterophilic* if it has a small homophily level, when most neighbors do not have the same label with source nodes. We aim to demonstrate that ACMP is a flexible GNN model which can learn well both kinds of datasets by balancing the attractive and repulsive forces. The GCN for examples cannot perform well for heterophilic dataset as its message passing aggregates only the neighbor (1-hop) nodes. The neural ODE is solved by `Torchdiffeq` package with Dormand–Prince adaptive step size scheme. Only few hyperparameters are needed to be tuned in our model. For all the experiments, we fine tune the learning rate, weight decay, dropout, hidden dimensional of the β which controls the repulsive force between nodes. We outline the details of hyperparameter search space in Appendix D.

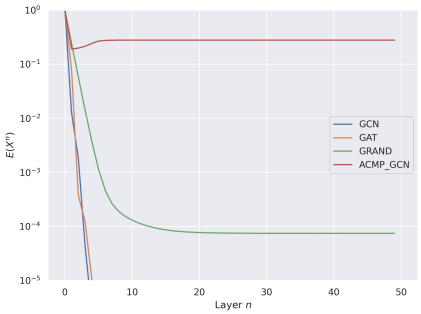

Figure 3: Evolution of Dirichlet energy $\mathbf{E}(\mathbf{X}^n)$ of layer-wise node features $\mathbf{X}^n$ propagated by GCN, GAT, GRAND, ACMP-GCN.

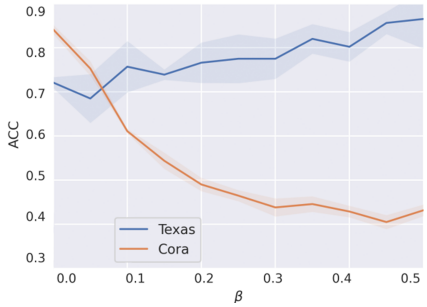

Figure 4: Significance plot for β in terms of test accuracy on Cora (orange) and Texas (blue) with 10 fixed random splits.

Homophilic datasets Our results are presented for the most widely used citation networks: Cora (McCallum et al., 2000), Citeseer (Sen et al., 2008) and Pubmed (Namata et al., 2012). Moreover, we evaluate our model on the Amazon co-purchasing graphs Computer and Photo (Namata et al., 2012), and CoauthorCS (Shchur et al., 2018). We compare our model with traditional GNN models: Graph Convolutional Network (GCN) (Kipf & Welling, 2017), Graph Attention Network (GAT) (Veličković et al., 2018), Mixture Model Networks (Monti et al., 2017) and GraphSage (Hamilton et al., 2017). We also compare our results with recent ODE-based GNNs, Continuous Graph Neural Networks (CGNN) (Xhonneux et al., 2020), Graph Neural Ordinary Differential Equations (GDE) (Poli et al., 2020) and Graph Neural Diffusion (GRAND) (Chamberlain et al., 2021). To address the limitations of this evaluation methodology proposed by Shchur et al. (2018), we report results for all datasets using 100 random splits with 10 random initialization's, and show the results in Table 1.

Table 1: Test accuracy and std for 10 initialization and 100 random train-val-test splits on six node classification benchmarks. **Red** (First), **blue** (Second), and **violet** (Third) are the best three methods.

Random Split Homophily level	**Cora** **0.83**	**CiteSeer** **0.71**	**PubMed** **0.79**	**Coauthor CS** **0.80**	**Computer** **0.77**	**Photo** **0.83**
GCN (Kipf & Welling, 2017)	81.5 ± 1.3	71.9 ± 1.9	77.8 ± 2.9	91.1 ± 0.5	82.6 ± 2.4	91.2 ± 1.2
GAT (Veličković et al., 2018)	81.8 ± 1.3	71.4 ± 1.9	78.7 ± 2.3	90.5 ± 0.6	78.0	85.7
GAT-ppr (Veličković et al., 2018)	81.6 ± 0.3	68.5 ± 0.2	76.7 ± 0.3	91.3 ± 0.1	$\mathbf{85.4 \pm 0.1}$	90.9 ± 0.3
MoNet (Monti et al., 2017)	81.3 ± 1.3	71.2 ± 2.0	78.6 ± 2.3	90.8 ± 0.6	83.5 ± 2.2	91.2 ± 2.3
GraphSage-mean (Hamilton et al., 2017)	79.2 ± 7.7	71.6 ± 2.0	77.4 ± 2.2	91.3 ± 2.8	82.4 ± 1.8	91.4 ± 1.3
GraphSage-max (Hamilton et al., 2017)	76.6 ± 1.9	67.5 ± 2.3	76.1 ± 2.3	85.0 ± 1.1	N/A	90.4 ± 1.3
CGNN (Xhonneux et al., 2020)	81.4 ± 1.6	66.9 ± 1.8	66.6 ± 4.4	$\mathbf{92.3 \pm 0.2}$	80.29 ± 2.0	91.39 ± 1.5
GDE (Poli et al., 2020)	78.7 ± 2.2	71.8 ± 1.1	73.9 ± 3.7	91.6 ± 0.1	81.9 ± 0.6	$\mathbf{92.4 \pm 2.0}$
GRAND-l (Chamberlain et al., 2021)	$\mathbf{83.6 \pm 1.0}$	$\mathbf{73.4 \pm 0.5}$	$\mathbf{78.8 \pm 1.7}$	$\mathbf{92.9 \pm 0.4}$	$\mathbf{83.7 \pm 1.2}$	$\mathbf{92.3 \pm 0.9}$
ACMP-GCN (ours)	$\mathbf{84.9 \pm 0.6}$	$\mathbf{75.0 \pm 1.0}$	$\mathbf{78.9 \pm 1.0}$	$\mathbf{93.0 \pm 0.5}$	83.5 ± 1.4	$\mathbf{91.8 \pm 1.1}$
ACMP-GAT (ours)	$\mathbf{82.3 \pm 0.5}$	$\mathbf{75.5 \pm 1.0}$	$\mathbf{79.4 \pm 0.4}$	91.8 ± 0.1	$\mathbf{84.4 \pm 1.6}$	91.1 ± 0.7

Heterophilic datasets We evaluate ACMP-GCN on the heterophilic graphs; Cornell, Texas and Wisconsin from the WebKB dataset[1]. In this case, the assumption of common neighbors does not hold. The poor performance of GCN and GAT models shown in Table 2 indicates that many GNN models struggle in this setting. Introducing repulsion can improve the performance of GNNs on heteroplilic datasets significantly. ACMP-GCN scores 30% higher than the original GCN for the Texas dataset which has the smallest homophily level among the datasets in the table.

Attractive and Repulsive interpretation As shown in Table 2 and Table 1, ACMP-GCN and ACMP-GAT achieve better performance than GCN and GAT on both homophilic and heterophilic datasets. The majority of $a_{i,j} - \beta$ in the homophilic are positive, which means most nodes are attracted to each other. Conversely, most $a_{i,j} - \beta$ for the heterophilic are negative, which means that most nodes are repelled by their neighbors. Several GNNs exploiting multi-hop information can

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/

Table 2: Node classification results on heterophilic datasets. We use the 10 fixed splits for training, validation and test from Pei et al. (2020) and show the mean and std of test accuracy. **Red** (First), **blue** (Second), and **violet** (Third) are the best three methods.

Homophily level	**Texas** 0.11	**Wisconsin** 0.21	**Cornell** 0.30
GPRGNN (Chien et al., 2021)	78.4 ± 4.4	82.9 ± 4.2	80.3 ± 8.1
H2GCN (Zhu et al., 2020)	84.9 ± 7.2	87.7 ± 5.0	82.7 ± 5.3
GCNII (Chen et al., 2020)	77.6 ± 3.8	80.4 ± 3.4	77.9 ± 3.8
Geom-GCN (Pei et al., 2020)	66.8 ± 2.7	64.5 ± 3.7	60.5 ± 3.7
PairNorm (Zhao & Akoglu, 2020)	60.3 ± 4.3	48.4 ± 6.1	58.9 ± 3.2
GraphSAGE (Hamilton et al., 2017)	82.4 ± 6.1	81.2 ± 5.6	76.0 ± 5.0
MLP	80.8 ± 4.8	85.3 ± 3.3	81.9 ± 6.4
GAT (Veličković et al., 2018)	52.2 ± 6.6	49.4 ± 4.1	61.9 ± 5.1
GCN (Kipf & Welling, 2017)	55.1 ± 5.2	51.8 ± 3.1	60.5 ± 5.3
GraphCON (Konstantin Rusch et al., 2022)	85.4 ± 4.2	87.8 ± 3.3	84.3 ± 4.8
ACMP-GCN (ours)	86.2 ± 3.0	86.1 ± 4.0	85.4 ± 7.0

achieve high performance in node classification (Zhu et al., 2020; Luan et al., 2021). However, high-order neighbor information will make the adjacency matrix dense and therefore can not be extended to large graphs, due to heavier computational cost. In our model, we take only one-hop information into account and add repulsive force ($\beta \geq 0$) to message passing, which has achieved the same or higher level of accuracy as multi-hop models in heterophilic datasets.

Performance of ACMP to β Hyperparameter β is a signal of the repulsive force, meaning that when $a_{ij} - \beta$ is negative, the two nodes repel one another. To illustrate β's impact, we use GCN as a diffusion term as a_{ij} do not change during the ODE process and all the changes are related to β. As shown by Figure 4, ACMP performs best in Cora (orange curve) when all nodes are attracted to one another i.e., all $a_{ij} - \beta$ is positive. As the beta increases, the performance of the model degrades. In contrast, for the Texas dataset, when all force is attractive, ACMP achieves only 70% accuracy (blue curve). As β increases, most $a_{ij} - \beta$ is negative, and the model's performance gets better. When all the force is repulsive, ACMP achieves highest accuracy on Texas datasets, which is in accordance with our claim that repulsive force is important for heterophilic datasets.

7 Related Work

Neural differential equations The topic of neural ODEs becomes an emerging field since E (2017) and Chen et al. (2018), with many follow-up works in the GNN field (Avelar et al., 2019; Poli et al., 2020; Sanchez-Gonzalez et al., 2019). GRAND (Chamberlain et al., 2021) propagated GNNs by the graph diffusion equation and Wu et al. (2023) developed an energy-constrained diffusion transformer. GraphCON (Konstantin Rusch et al., 2022) employed a second-order system to conquer oversmoothing of deep graph neural networks. By exploiting the fixed point of the dynamical system, Gallicchio & Micheli (2020) proposed FDGNN as an approach to graph classification.

Allen-Cahn based variational graph models In Bertozzi & Flenner (2012); Luo & Bertozzi (2017); Merkurjev et al. (2013) and references therein, authors extended Allen-Cahn related potential to graphical framework and developed a class of variational algorithms to solve the clustering, semisupervised learning and graph cutting problems. The new ingredient of graph neural network which enables us to combine learnable attraction and repulsion separates our method from the classical variational graph models.

8 Conclusion

We develop a new message passing method with simple implementation. The method is based on the Allen-Cahn particle system with repulsive force. The proposed ACMP inherits the characteristic dynamics of the particle system and thus shows adaption for node classification tasks with high homophily difficulty. Also, it propels networks to dozens of layers without getting oversmoothing. A strictly positive lower bound of the Dirichlet energy is shown by theoretical and experimental

results which guarantees non-oversmoothing of ACMP. Experiments show excellent performance of the model for various real datasets.

REFERENCES

Giacomo Albi, Nicola Bellomo, Luisa Fermo, S-Y Ha, J Kim, Lorenzo Pareschi, David Poyato, and Juan Soler. Vehicular traffic, crowds, and swarms: From kinetic theory and multiscale methods to applications and research perspectives. *Mathematical Models and Methods in Applied Sciences*, 29(10):1901–2005, 2019.

Samuel M Allen and John W Cahn. A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica*, 27(6):1085–1095, 1979.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *ICLR*, 2021.

Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, pp. 1–10, 2021.

Pedro HC Avelar, Anderson R Tavares, Marco Gori, and Luis C Lamb. Discrete and continuous deep residual learning over graphs. *arXiv preprint arXiv:1911.09554*, 2019.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a 3-track network. *Science*, 373:871–876, 2021.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çaglar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matthew Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018.

Andrea L Bertozzi and Arjuna Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik Bekkers, and Max Welling. Geometric and physical quantities improve E(3) equivariant message passing. In *ICLR*, 2022a.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *ICLR*, 2022b. URL https://openreview.net/forum?id=vSix3HPYKSU.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. In *ICML Workshop on Graph Representation Learning*, 2020.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.

Benjamin Paul Chamberlain, James Rowbottom, Maria I. Gorinova, Stefan D Webb, Emanuele Rossi, and Michael M. Bronstein. GRAND: Graph neural diffusion. In *ICML*, 2021. URL https://openreview.net/forum?id=_1fu_cjsaRE.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *ICML*, pp. 1725–1735. PMLR, 2020.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, volume 31, 2018.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *ICLR*, 2021. URL https://openreview.net/forum?id=n6jl7fLxrP.

Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007.

Pierre Degond and Sébastien Motsch. Large scale dynamics of the persistent turning walker model of fish behavior. *Journal of Statistical Physics*, 131(6):989–1021, 2008.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

Di Fang, Seung-Yeal Ha, and Shi Jin. Emergent behaviors of the Cucker-Smale ensemble under attractive-repulsive couplings and rayleigh frictions. *Mathematical Models and Methods in Applied Sciences*, 29(07):1349–1385, 2019.

Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, D Boscaini, MM Bronstein, and BE Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Claudio Gallicchio and Alessio Micheli. Fast and deep graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3898–3905, 2020.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *NeurIPS*, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

Seung-Yeal Ha and Eitan Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinetic & Related Models*, 1(3):415, 2008.

Seung-Yeal Ha, Taeyoung Ha, and Jong-Ho Kim. Asymptotic dynamics for the Cucker-Smale-type model with the Rayleigh friction. *Journal of Physics A: Mathematical and Theoretical*, 43(31): 315201, 2010.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

T. Konstantin Rusch, Benjamin P. Chamberlain, James Rowbottom, Siddhartha Mishra, and Michael M. Bronstein. Graph-coupled oscillator networks. In *ICML*, 2022.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in Neural Information Processing Systems*, 34:20887–20902, 2021.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *ICML*, pp. 3276–3285, 2018.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Is heterophily a real nightmare for graph neural networks to do node classification? *arXiv preprint arXiv:2109.05641*, 2021.

Xiyang Luo and Andrea L Bertozzi. Convergence of the graph Allen-Cahn scheme. *Journal of Statistical Physics*, 167(3):934–958, 2017.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Ekaterina Merkurjev, Tijana Kostic, and Andrea L Bertozzi. An MBO scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.

Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *CVPR*, pp. 5115–5124, 2017.

Sebastien Motsch and Eitan Tadmor. A new model for self-organized dynamics and its flocking behavior. *Journal of Statistical Physics*, 144(5):923–947, 2011.

Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014.

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8, pp. 1, 2012.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. On second order behaviour in augmented neural ODEs. In *NeurIPS*, volume 33, pp. 5911–5921, 2020.

Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *ICLR*, 2019.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: geometric graph convolutional networks. In *ICLR*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. Graph neural ordinary differential equations. In *AAAI Workshop on Deep Learning on Graphs: Methodologies and Applications*, 2020.

Anton V Proskurnikov and Roberto Tempo. A tutorial on modeling and analysis of dynamic social networks. Part I. *Annual Reviews in Control*, 43:65–79, 2017.

Loan Rayleigh. *JWS, The Theory of Sound, vol. 1*. Macmillan, London (reprinted Dover, New York, 1945), 1894.

Alvaro Sanchez-Gonzalez, Victor Bapst, Kyle Cranmer, and Peter Battaglia. Hamiltonian graph networks with ODE integrators. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2019.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–93, 2008.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, and Junchi Yan. Difformer: Scalable (graph) transformers induced by energy constrained diffusion. *arXiv preprint arXiv:2301.09474*, 2023.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020.

Louis-Pascal Xhonneux, Meng Qu, and Jian Tang. Continuous graph neural networks. In *ICML*, pp. 10432–10441, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2018.

Yujun Yan, Milad Hashemi, Kevin Swersky, Yaoqing Yang, and Danai Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. *arXiv preprint arXiv:2102.06462*, 2021.

Lingxiao Zhao and Leman Akoglu. PairNorm: Tackling oversmoothing in GNNs. In *ICLR*, 2020. URL https://openreview.net/forum?id=rkecl1rtwB.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *NeurIPS*, volume 33, pp. 7793–7804, 2020.