Graph Neural Networks with Adaptive Residual

Xiaorui Liu

Joint work with Jiayuan Ding, Wei Jin,

Han Xu, Yao Ma, Zitao Liu, Jiliang Tang

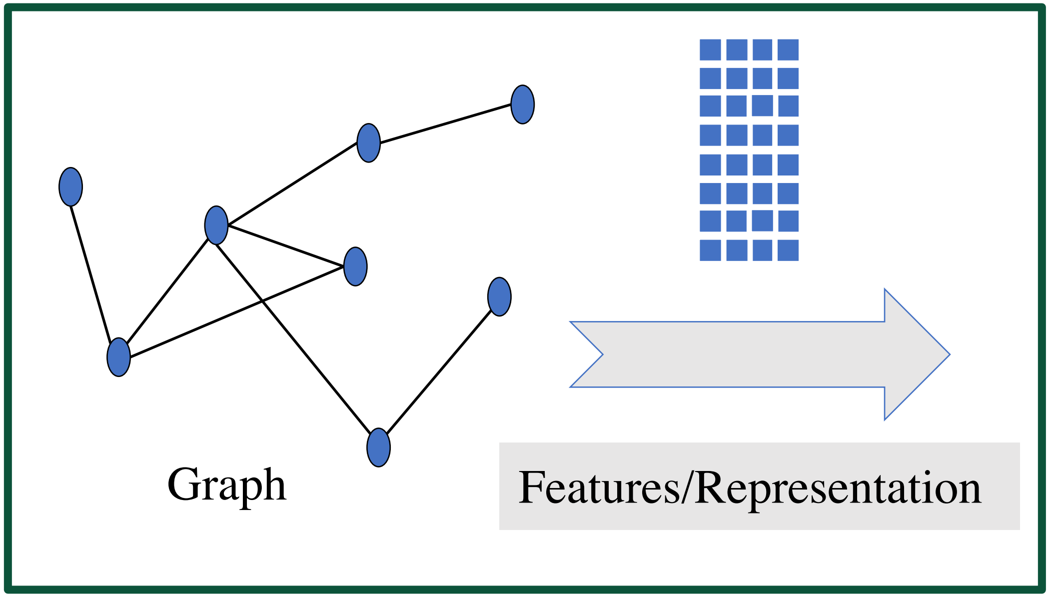
Michigan State University

New Jersey Institute of Technology

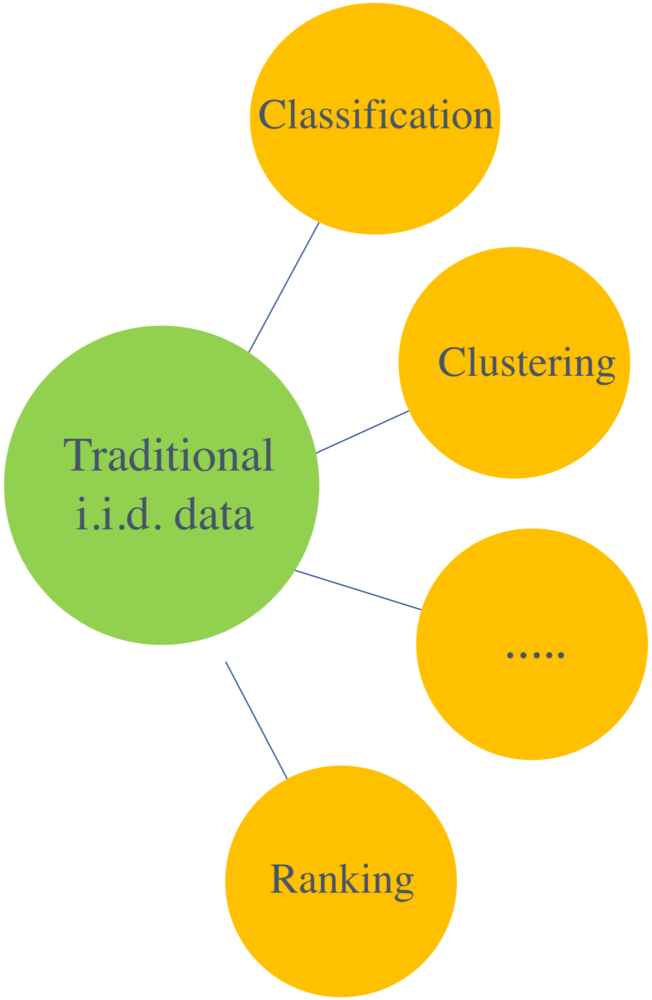
TAL Education Group

NeurIPS 2021, December

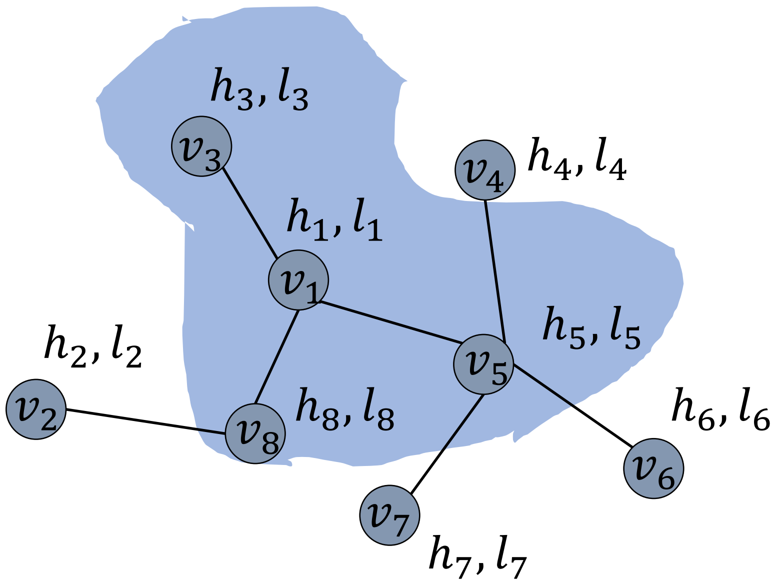
Machine Learning on Graphs



Representation Learning on Graphs



Graph Neural Networks



Message Passing

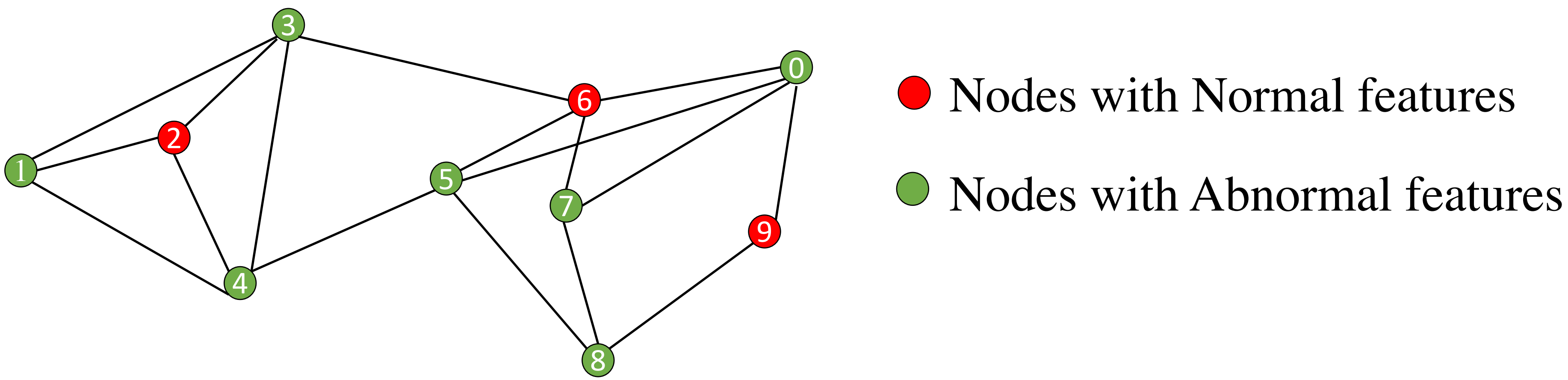
$$m_i^{(k+1)} = \sum_{v_j \in N(v_i)} M_k\left(h_i^{(k)}, h_j^{(k)}, e_{ij}\right)$$

Feature Updating

$$h_i^{(k+1)} = U_k\left(h_i^{(k)}, m_i^{(k+1)}\right)$$

Neural Message Passing for Quantum Chemistry, Justin Gilmer et al**,** ICML 2017

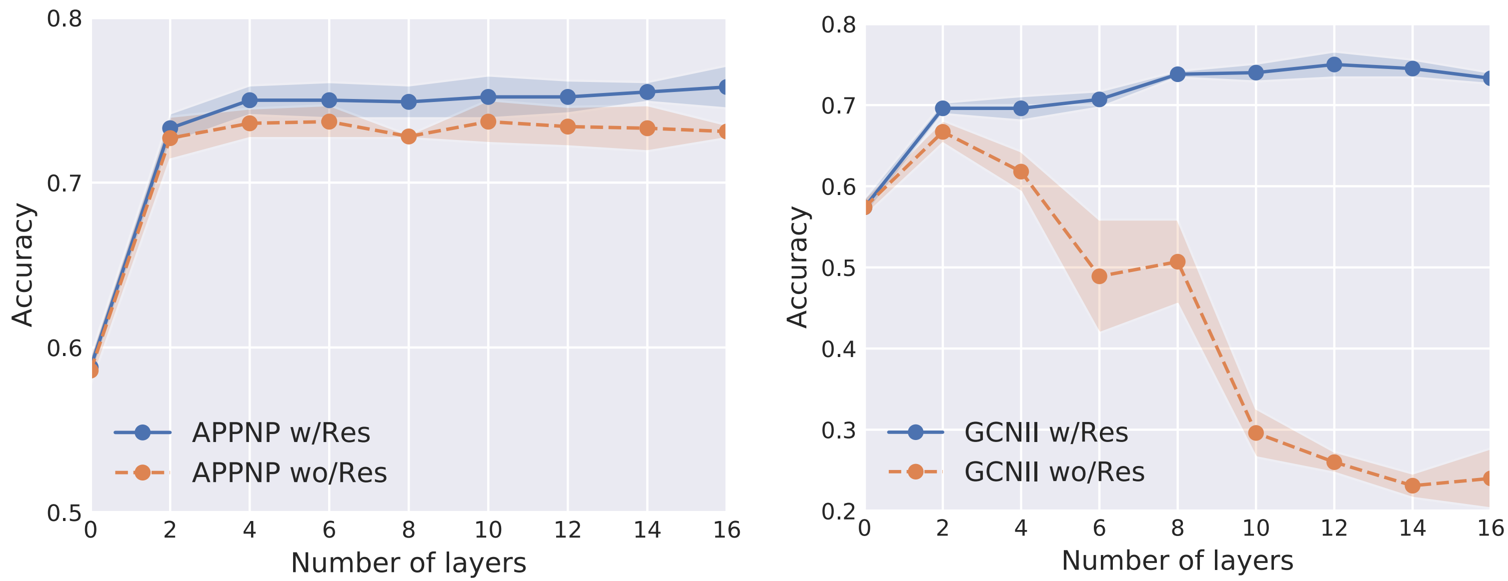
A Practical Scenario



Examples of abnormal features

- **Missing feature**: new users in social networks
- **Noisy feature**: uncertainty and dynamics in traffic information
- **Adversarial features**: node attributes are maliciously manipulated

An Interesting Discovery



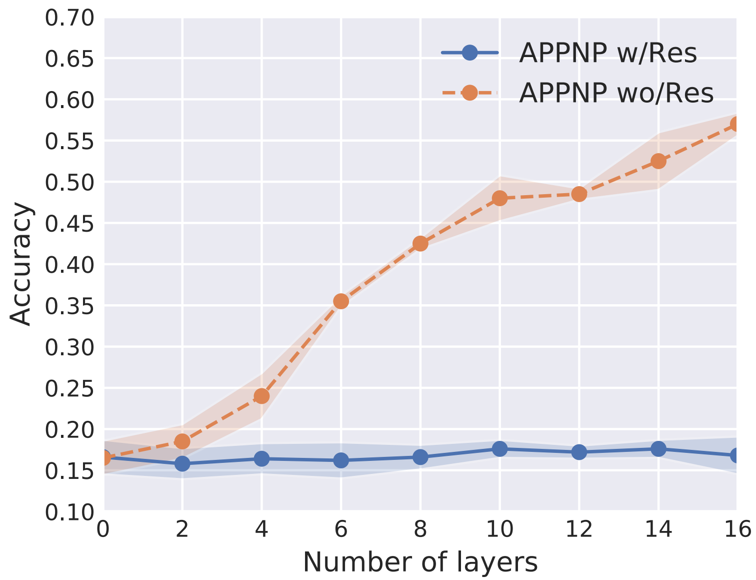
Node classification accuracy for nodes with normal features
(with 10% noisy nodes in Cora)

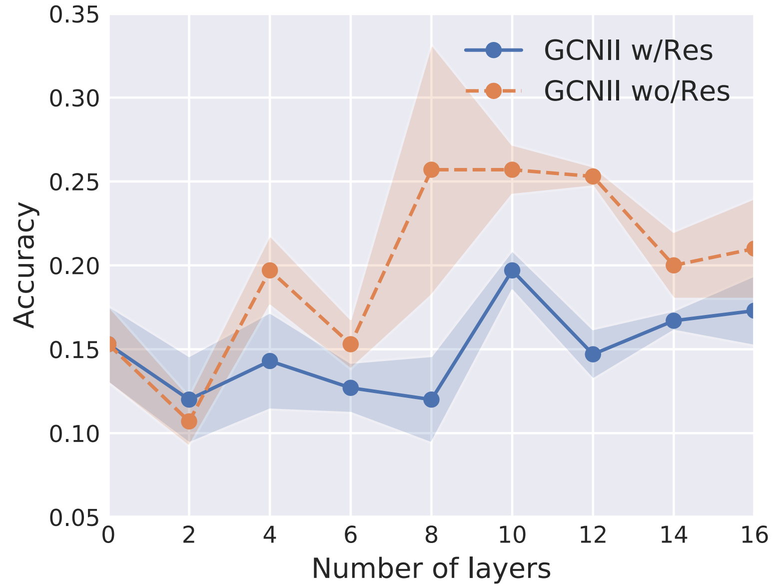
Finding I: (necessity of residual)

(1) Residual connection helps GNNs benefit from more layers;

(2) Without residual, too many aggregations could hurt the performance.

An Interesting Discovery (Cont.)





Node classification accuracy for nodes with noisy features
(with 10% noisy nodes in Cora)

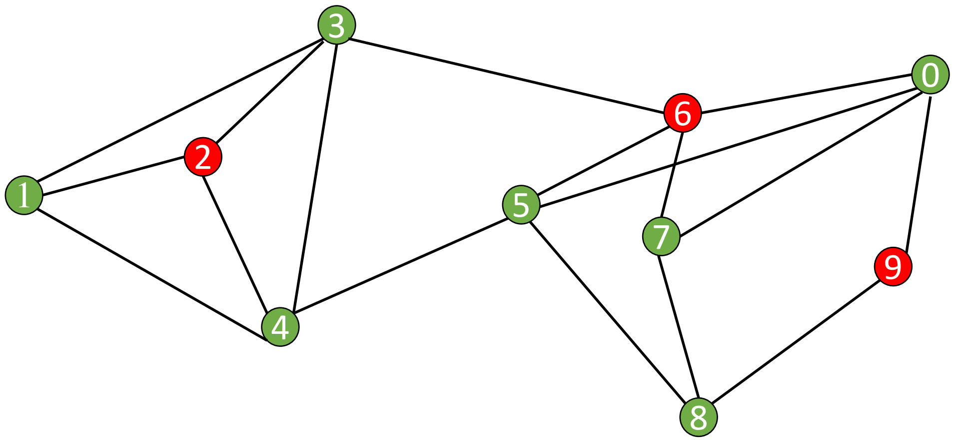
Finding II: (necessity of aggregation)

(1) Feature aggregations can boost the performance for noisy nodes;

(2) Residual connection makes GNNs more fragile to noisy node features.

An Interesting Discovery (Cont.)

A practical scenario



Nodes with Normal features

Nodes with Abnormal features

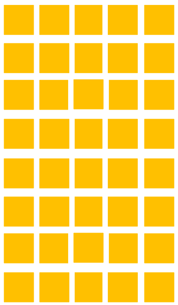
The dilemma:

(1) Normal features need residual connections to avoid over-smoothing

(2) Residual connections hurt the performance for nodes with abnormal features

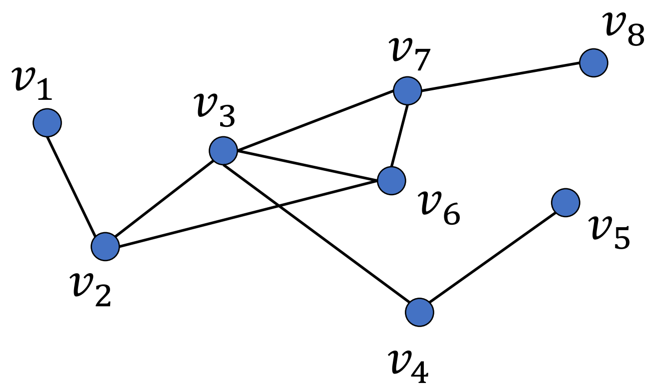
Understanding from GSP

"Noisy Signal"

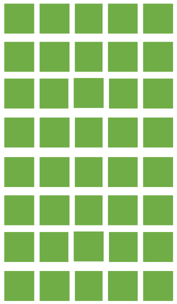


$\mathbf{X}_{\mathrm{in}}$

Graph



"Clean Signal"



$\mathbf{F}$

"Nodes are similar as their neighbors"

$$\arg\min_{\mathbf{X}\in\mathbb{R}^{n\times d}} \mathcal{L}(\mathbf{X}) := \frac{\alpha}{2(1-\alpha)}\|\mathbf{X}-\mathbf{X}_{\mathrm{in}}\|_F^2 + \frac{1}{2}\mathrm{tr}\Big(\mathbf{X}^\top(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{X}\Big)$$

Proximity to the input **Smoothness prior**

A unified view on graph neural networks as graph signal denoising, Yao Ma, Xiaorui Liu et al, 2020

Understanding from GSP

$$\arg\min_{\mathbf{X}\in\mathbb{R}^{n\times d}} \mathcal{L}(\mathbf{X}) := \frac{\alpha}{2(1-\alpha)}\|\mathbf{X}-\mathbf{X}_{\text{in}}\|_F^2 + \frac{1}{2}\operatorname{tr}\left(\mathbf{X}^\top(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{X}\right)$$

Proximity to the input **Smoothness prior**

Define Prior ⇨ Optimization Solver ⇨ Message Passing

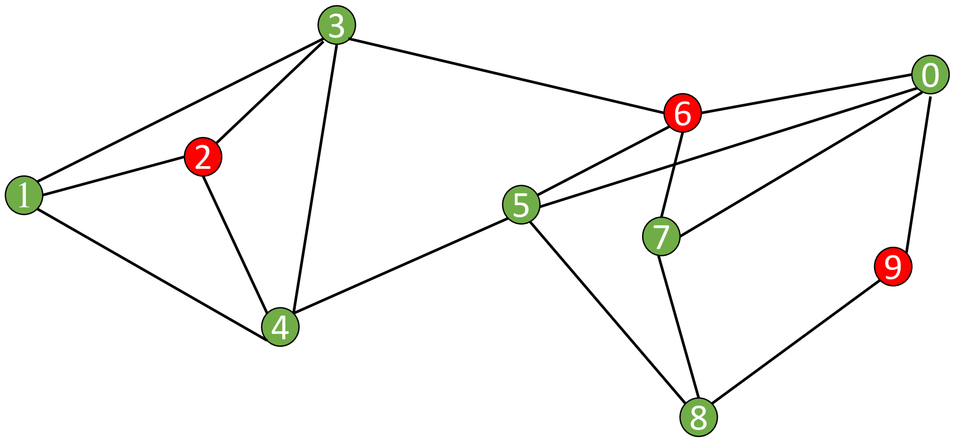
GCN $\mathbf{X}_{\text{out}} = \tilde{\mathbf{A}}\mathbf{X}_{\text{in}}$ APPNP/GCNII $\mathbf{X}^{(k+1)} = (1-\alpha)\tilde{\mathbf{A}}\mathbf{X}^{(k)} + \alpha\mathbf{X}_{\text{in}}$

- **Feature aggregation**: correct abnormal features by smoothing
- **Residual connection**: reduce feature over-smoothing by maintaining feature proximity but carry undesirable abnormal features

A unified view on graph neural networks as graph signal denoising, Yao Ma, Xiaorui Liu et al, 2020

Motivation

A practical scenario



- Nodes with Normal features
- Nodes with Abnormal features

Can we design a message passing with node-wise adaptive feature aggregation and residual connection to achieve good performance on both types of nodes?

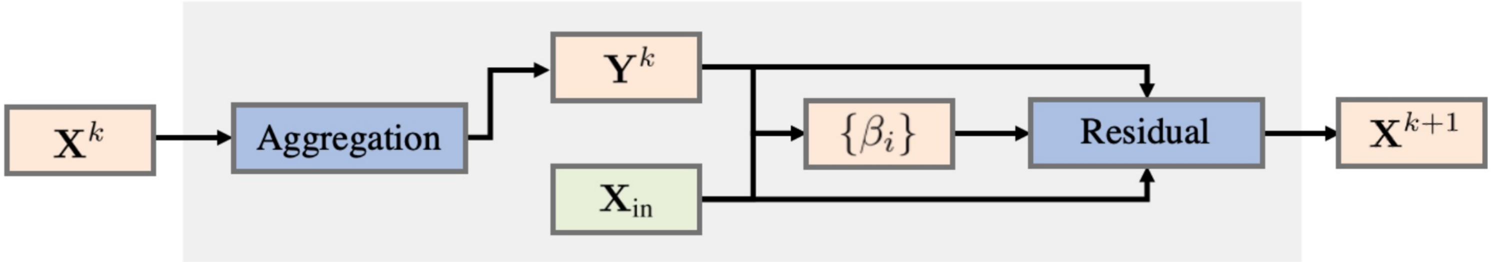
Motivation

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times d}}{\arg\min} \quad \mathcal{L}(\mathbf{X}) := \lambda\|\mathbf{X}-\mathbf{X}_{\text{in}}\|_{21} + (1-\lambda)\mathrm{tr}(\mathbf{X}^\top(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{X})$$

$$\|\mathbf{X}-\mathbf{X}_{\text{in}}\|_{21} := \sum_{i=1}^{n}\|\mathbf{X}_i-(\mathbf{X}_{\text{in}})_i\|_2$$

- Laplacian smoothing with the robust feature proximity
- Tolerate large deviations due to the less aggressive penalty on large values
- Potential removal of abnormal features

Adaptive Message Passing



$$\begin{cases} \mathbf{Y}^k & = \big(1-2\gamma(1-\lambda)\big)\mathbf{X}^k + 2\gamma(1-\lambda)\tilde{\mathbf{A}}\mathbf{X}^k \\ \beta_i & = \max(1-\dfrac{\gamma\lambda}{\|\mathbf{Y}_i^k-(\mathbf{X}_{\text{in}})_i\|_2},0) \quad \forall i\in[n] \\ \mathbf{X}_i^{k+1} & = (1-\beta_i)(\mathbf{X}_{\text{in}})_i+\beta_i\mathbf{Y}_i^k \qquad \forall i\in[n] \end{cases}$$

Node-wise adaptive message passing

Adaptive Message Passing

$$\begin{cases} \mathbf{Y}^k &= \big(1 - 2\gamma(1-\lambda)\big)\mathbf{X}^k + 2\gamma(1-\lambda)\tilde{\mathbf{A}}\mathbf{X}^k \\ \beta_i &= \max(1 - \dfrac{\gamma\lambda}{\|\mathbf{Y}_i^k - (\mathbf{X}_{\text{in}})_i\|_2}, 0) \quad \forall i \in [n] \\ \mathbf{X}_i^{k+1} &= (1-\beta_i)(\mathbf{X}_{\text{in}})_i + \beta_i \mathbf{Y}_i^k \qquad \forall i \in [n] \end{cases}$$

Node-wise adaptive message passing

Interpretation as feature selection

- Small residual ($\beta_i \to 1$): if $(\mathbf{X}_{\text{in}})_\text{i}$ is significantly inconsistent with local neighbors;
- Large residual ($\beta_i \to 0$): if $(\mathbf{X}_{\text{in}})_\text{i}$ is very consistent with local neighbors;
- β_i provides a natural transition from 0 to 1 modulated by λ

AirGNN

- Adaptive message passing (AMP) can be used as a building block in many GNN architecture

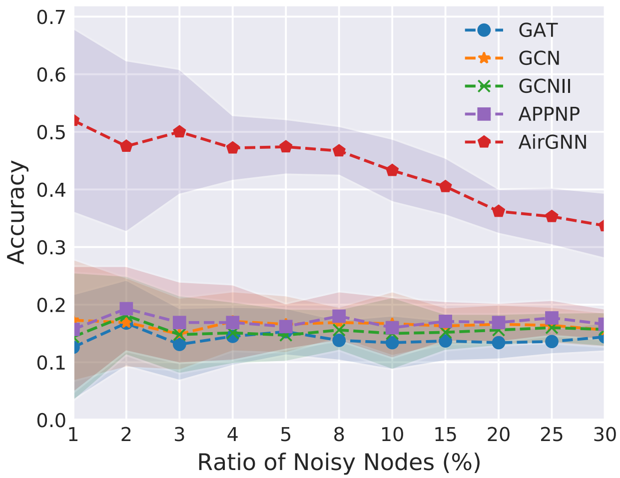
- In this work, we follow the decoupled architecture as APPNP

$$\mathbf{X}_{\text{in}} = h_\theta(\mathbf{X}_{\text{fea}})$$
$$\mathbf{Y}_{\text{pre}} = \mathbf{AMP}\left(\mathbf{X}_{\text{in}}, K, \lambda\right)$$

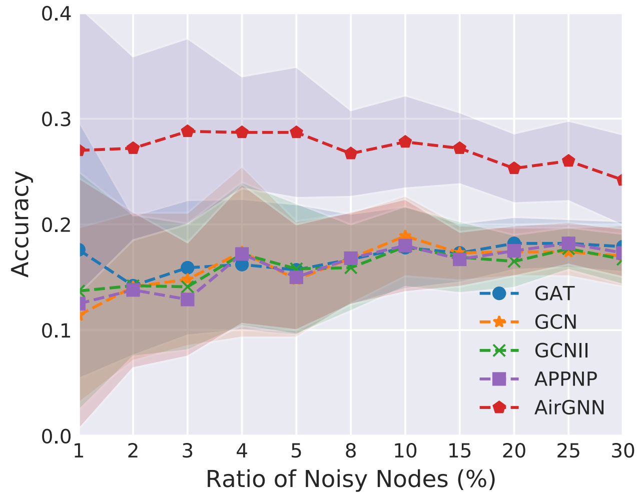
- Parameters θ are trained by the cross-entropy loss defined on labeled data through back propagation

AirGNN: GNN with Adaptive Residual

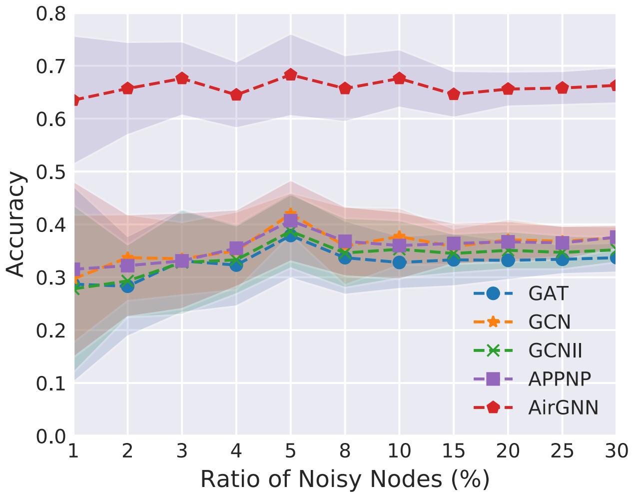
Experiment in Noise Setting



Cora



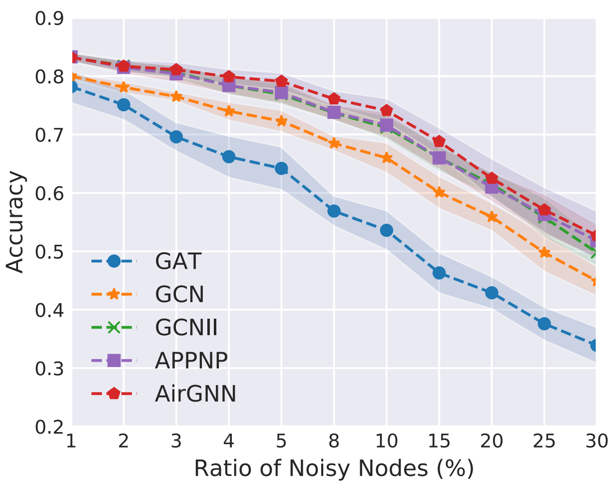
Citeseer



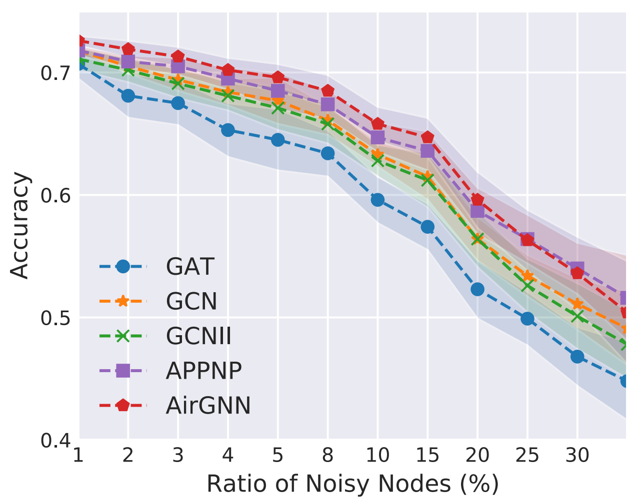
PubMed

Node classification performance on nodes with noisy features

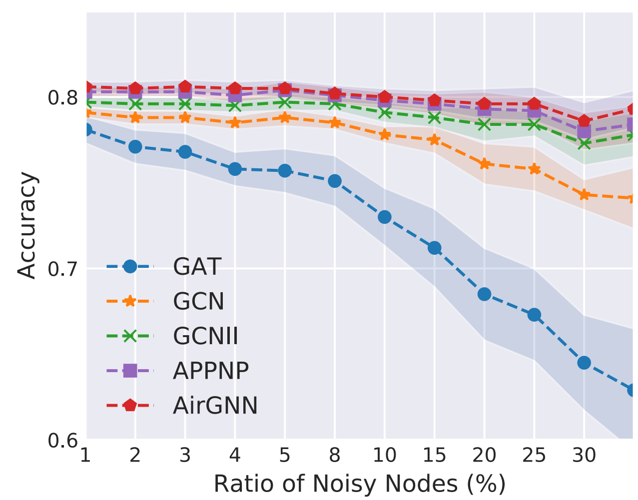
Experiment in Noise Settings



Cora



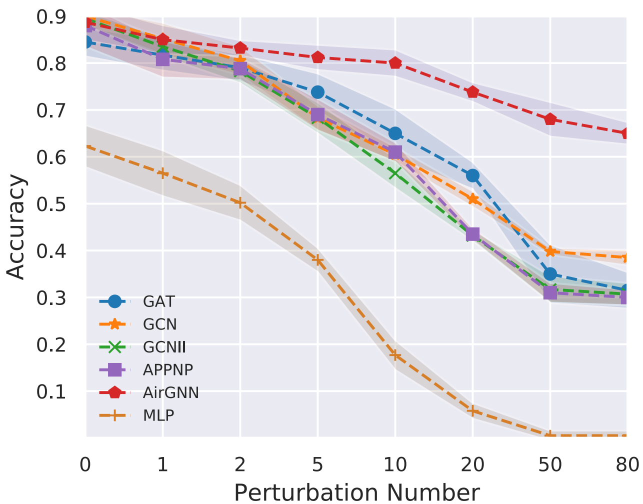
Citeseer



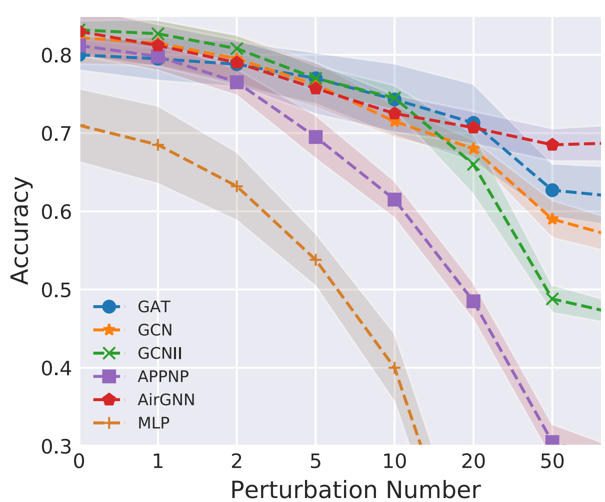
PubMed

Node classification performance on normal nodes

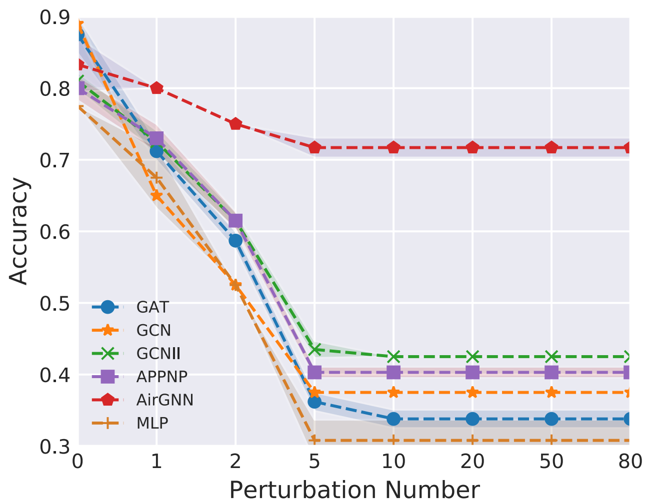
Experiment in Adversarial Settings



Cora



Citeseer



PubMed

Node classification performance on adversarially perturbed nodes

Adaptive Residual

Table 1: Average adaptive score (β) and residual weight ($1-\beta$) in the noisy feature scenario.

Measure	**Cora**	**CiteSeer**	**PubMed**
Average adaptive score for abnormal nodes	0.998 ± 0.000	0.988 ± 0.000	0.996 ± 0.000
Average adaptive score for normal nodes	0.924 ± 0.002	0.807 ± 0.005	0.869 ± 0.006
Average residual weight for abnormal nodes	0.002 ± 0.000	0.012 ± 0.000	0.004 ± 0.000
Average residual weight for normal nodes	0.076 ± 0.002	0.193 ± 0.005	0.131 ± 0.006

Table 2: Average adaptive score (β) and residual weight ($1-\beta$) in the adversarial feature scenario.

Measure	**Cora**	**CiteSeer**	**PubMed**
Average adaptive score for abnormal nodes	0.987 ± 0.000	0.930 ± 0.007	0.959 ± 0.005
Average adaptive score for normal nodes	0.922 ± 0.004	0.689 ± 0.024	0.826 ± 0.016
Average residual weight for abnormal nodes	0.013 ± 0.000	0.070 ± 0.007	0.041 ± 0.005
Average residual weight for normal nodes	0.078 ± 0.004	0.311 ± 0.024	0.174 ± 0.016

Conclusion

Summary

- Discover the intrinsic tension between feature aggregation and residual connection in GNNs
- Design a simple and effective adaptive message passing scheme that can be used a building block to improve the robustness against abnormal features
- Design AirGNN that achieves impressive performance improvement in multiple abnormal settings
- Verify that the adaptive residual is a good indicator of abnormal features in ablation study

Code: https://github.com/lxiaorui/AirGNN