Diverse Message Passing

Liang Yang
About Me

2000 |
2004 |
2007 |
2007 |
2009 |
2009 |
2010 |
2010 |
2013 |
2016 |
2018 |
2018 |
today

Homepage: http://yangliang.github.io/
Outline

• Existing Message Passing Framework

• Diverse Message Passing
  • Motivations
  • Semi-supervised Task
  • Self-supervised Task

• Theoretical Analysis

• Conclusions
Existing Message Passing

Aggregation-Combination

\[
\bar{h}_v^k = \text{AGGREGATE}^k \left( \{ h_u^{k-1} | u \in \mathcal{N}(v) \} \right),
\]

\[
h_v^k = \text{COMBINE}^k \left( h_v^{k-1}, \bar{h}_v^k \right),
\]

\[
h_v^k = \sigma \left( \left( c_{vv}^k h_v^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k h_u^{k-1} \right) W^k \right),
\]

Uniform

In different attribute channels
Diverse Message Passing

(a) Uniform and Blind Message Passing

(b) Diverse and Interactive Message Passing
Motivations - Attribute Diversity

Homophily Rate of Graph

$$\beta = \frac{1}{N} \sum_{v \in V} \frac{\text{Number of } v\text{'s neighbors who have the same label as } v}{\text{Number of } v\text{'s neighbors}}.$$ 

Homophily Rates of Attributes

$$\beta_f = \frac{1}{\sum_{v \in V} x_{vf}} \sum_{v \in V} \beta_{vf} = \frac{1}{\sum_{v \in V} x_{vf}} \sum_{v \in V} \left( \frac{x_{vf} \sum_{u \in N(v)} x_{uf}}{d_v} \right),$$

Graph-level

- Topology + Node Label

Attribute-level

- Topology + Node Attribute

Node Attribute

- Texas (0.06)
- Wisconsin (0.16)
- Chameleon (0.25)
- Cornell (0.11)
- Actor (0.24)
- Citeseer (0.71)
- Pubmed (0.79)
- Cora (0.83)
Semi-supervised Diverse Message Passing

\( h_v^k = \sigma \left( \left( c_{vv}^k h_v^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k h_u^{k-1} \right) w^k \right) \),

\( h_v^k = \sigma \left( \left( c_{vv}^k \oplus h_v^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k \oplus h_u^{k-1} \right) w^k \right) \).
Semi-supervised Learning

The first strategy

\[ c_{uv}^k = \tanh \left( [h_{v}^{k-1} || h_{u}^{k-1}] W_{c}^k \right), \]

Model Complexity \( O\left( F \times F \right) \)

The second strategy

\[ c_{v}^k = \tanh \left( [h_{v}^{k-1} || \bar{h}_{v}^{k-1}] W_{c}^k \right), \]

\[ \bar{h}_{v}^{k} = \text{AGGREGATE}^{k} \left( \{ h_{u}^{k-1} | u \in \mathcal{N}(v) \} \right), \]

Model Complexity \( O\left( F \times F \right) \)
<table>
<thead>
<tr>
<th>Methods</th>
<th>Texas</th>
<th>Wisconsin</th>
<th>Actor</th>
<th>Squirrel</th>
<th>Cham.</th>
<th>Cornell</th>
<th>Citeseer</th>
<th>Pubmed</th>
<th>Cora</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphSAGE</td>
<td>82.43</td>
<td>81.18</td>
<td>34.23</td>
<td>41.61</td>
<td>58.73</td>
<td>75.95</td>
<td>76.04</td>
<td>88.45</td>
<td>86.90</td>
</tr>
<tr>
<td>GCN</td>
<td>64.86</td>
<td>56.86</td>
<td>31.12</td>
<td>32.28</td>
<td>53.51</td>
<td>54.05</td>
<td>75.53</td>
<td>84.71</td>
<td>85.51</td>
</tr>
<tr>
<td>GAT</td>
<td>58.38</td>
<td>55.29</td>
<td>26.28</td>
<td>30.62</td>
<td>54.69</td>
<td>58.92</td>
<td>75.46</td>
<td>84.68</td>
<td>82.68</td>
</tr>
<tr>
<td>SAGE+JK</td>
<td>83.78</td>
<td>81.96</td>
<td>34.28</td>
<td>40.85</td>
<td>58.11</td>
<td>75.68</td>
<td>76.05</td>
<td>88.34</td>
<td>85.96</td>
</tr>
<tr>
<td>Cheby+JK</td>
<td>78.38</td>
<td>82.55</td>
<td>35.14</td>
<td>45.03</td>
<td><strong>63.79</strong></td>
<td>74.59</td>
<td>74.98</td>
<td>89.07</td>
<td>85.49</td>
</tr>
<tr>
<td>GCN+JK</td>
<td>66.49</td>
<td>74.31</td>
<td>34.18</td>
<td>40.45</td>
<td>63.42</td>
<td>64.59</td>
<td>74.51</td>
<td>88.41</td>
<td>85.79</td>
</tr>
<tr>
<td>GCN-Cheby</td>
<td>77.30</td>
<td>79.41</td>
<td>34.11</td>
<td>43.86</td>
<td>55.24</td>
<td>74.32</td>
<td>75.82</td>
<td>88.72</td>
<td>86.76</td>
</tr>
<tr>
<td>MixHop</td>
<td>77.84</td>
<td>75.88</td>
<td>32.22</td>
<td>43.80</td>
<td>60.50</td>
<td>73.51</td>
<td>76.26</td>
<td>85.31</td>
<td><strong>87.61</strong></td>
</tr>
<tr>
<td>GEOM-GCN</td>
<td>67.57</td>
<td>64.12</td>
<td>31.63</td>
<td>38.14</td>
<td>60.90</td>
<td>60.81</td>
<td><strong>77.99</strong></td>
<td>90.05</td>
<td>85.27</td>
</tr>
<tr>
<td>H2GCN</td>
<td>84.86</td>
<td>86.67</td>
<td><strong>35.86</strong></td>
<td>36.42</td>
<td>57.11</td>
<td>82.16</td>
<td>77.04</td>
<td>89.40</td>
<td>86.92</td>
</tr>
<tr>
<td>DMP-Deg</td>
<td>78.38</td>
<td>80.39</td>
<td>33.09</td>
<td>32.46</td>
<td>54.38</td>
<td>83.78</td>
<td>76.87</td>
<td>88.10</td>
<td>86.31</td>
</tr>
<tr>
<td>DMP-2-Sum</td>
<td>78.37</td>
<td>84.31</td>
<td>34.93</td>
<td>32.18</td>
<td>55.92</td>
<td>83.78</td>
<td>76.27</td>
<td>88.15</td>
<td>85.31</td>
</tr>
<tr>
<td>DMP-2-Con</td>
<td>83.78</td>
<td>84.31</td>
<td>34.67</td>
<td>44.28</td>
<td>60.53</td>
<td>83.78</td>
<td>75.97</td>
<td>85.31</td>
<td>85.31</td>
</tr>
<tr>
<td>DMP-1-Posi</td>
<td>86.48</td>
<td>84.31</td>
<td>35.72</td>
<td>34.96</td>
<td>51.53</td>
<td>70.27</td>
<td>75.67</td>
<td>88.10</td>
<td>86.11</td>
</tr>
<tr>
<td>DMP-1-Sum</td>
<td>86.48</td>
<td>86.27</td>
<td>34.21</td>
<td>43.42</td>
<td>50.21</td>
<td>70.27</td>
<td>76.13</td>
<td>88.13</td>
<td>82.28</td>
</tr>
<tr>
<td>DMP-1-Con</td>
<td><strong>89.19</strong></td>
<td><strong>92.16</strong></td>
<td>35.06</td>
<td><strong>47.26</strong></td>
<td>62.28</td>
<td><strong>89.19</strong></td>
<td>76.43</td>
<td>89.27</td>
<td>86.52</td>
</tr>
</tbody>
</table>
Figure 2: Classification accuracy results with various depths.

Figure 3: Distributions of learned weights of sampled attribute dimensions.
Self-supervised Diverse Message Passing

Reduce Model Complexity & Preserve Expressive Power

\[ h^k_v = \sigma \left( \left( c^k_{vv} \odot h^{k-1}_v \right) + \sum_{u \in \mathcal{N}(v)} c^k_{uv} \odot h^{k-1}_u \right) w^k \right) \]

\[ h^k_v = \sigma \left( m^k_{vv} + \sum_{u \in \mathcal{N}(v)} m^k_{uv} w^k \right) \]

**Message**

\[ m^k_{uv} = \frac{h^{k-1}_v \odot h^{k-1}_u}{\langle h^{k-1}_v, h^{k-1}_u \rangle} \]

Diverse Interactive
### Table 3: Node Classification Results in Terms of Accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cora</th>
<th>CiteSeer</th>
<th>PubMed</th>
<th>Amazon-C</th>
<th>Amazon-P</th>
<th>Coauthor-CS</th>
<th>Coauthor-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>58.2±2.1</td>
<td>59.1±2.3</td>
<td>70.0±2.1</td>
<td>44.9±5.8</td>
<td>69.6±3.8</td>
<td>88.3±0.7</td>
<td>88.9±1.1</td>
</tr>
<tr>
<td>LogReg</td>
<td>57.1±2.3</td>
<td>61.0±2.2</td>
<td>64.1±3.1</td>
<td>64.1±5.7</td>
<td>73.0±6.5</td>
<td>86.4±0.9</td>
<td>86.7±1.5</td>
</tr>
<tr>
<td>LP</td>
<td>68.0±0.2</td>
<td>45.3±0.2</td>
<td>63.0±0.5</td>
<td>70.8±0.0</td>
<td>67.8±0.0</td>
<td>74.3±0.0</td>
<td>90.2±0.2</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>81.2±0.5</td>
<td>69.8±0.5</td>
<td>74.4±0.3</td>
<td>62.6±0.0</td>
<td>74.3±0.0</td>
<td>91.5±0.0</td>
<td>92.1±0.3</td>
</tr>
<tr>
<td>GCN</td>
<td>81.5±0.2</td>
<td>70.3±0.3</td>
<td>79.0±0.4</td>
<td>76.3±0.5</td>
<td>87.3±1.0</td>
<td><strong>91.8±0.1</strong></td>
<td><strong>92.6±0.7</strong></td>
</tr>
<tr>
<td>GAT</td>
<td><strong>83.0±0.7</strong></td>
<td><strong>72.5±0.7</strong></td>
<td><strong>79.0±0.3</strong></td>
<td>79.3±1.1</td>
<td>86.2±1.5</td>
<td>90.5±0.7</td>
<td>91.3±0.6</td>
</tr>
<tr>
<td>MoNet</td>
<td>81.3±1.3</td>
<td>71.2±2.0</td>
<td>78.6±2.3</td>
<td>83.5±2.2</td>
<td><strong>91.2±1.3</strong></td>
<td>90.8±0.6</td>
<td>92.5±0.9</td>
</tr>
<tr>
<td>DGI</td>
<td>81.7±0.6</td>
<td>71.5±0.7</td>
<td>77.3±0.6</td>
<td>75.9±0.6</td>
<td>83.1±0.5</td>
<td>90.0±0.3</td>
<td>91.3±0.4</td>
</tr>
<tr>
<td>GMI</td>
<td>80.9±0.7</td>
<td>71.1±0.3</td>
<td>77.0±0.2</td>
<td>76.8±0.1</td>
<td>85.1±0.1</td>
<td>90.1±0.0</td>
<td>OOM</td>
</tr>
<tr>
<td>MVGRL</td>
<td>82.9±0.7</td>
<td>72.6±0.7</td>
<td>79.4±0.3</td>
<td>79.0±0.6</td>
<td>87.3±0.3</td>
<td>88.4±0.3</td>
<td>92.6±0.4</td>
</tr>
<tr>
<td>GRACE</td>
<td>80.0±0.4</td>
<td>71.7±0.6</td>
<td>79.5±1.1</td>
<td>71.8±0.4</td>
<td>81.8±1.0</td>
<td>90.1±0.8</td>
<td>92.3±0.6</td>
</tr>
<tr>
<td>GCA</td>
<td>81.0±0.4</td>
<td>71.9±0.5</td>
<td>80.5±1.1</td>
<td>80.8±0.4</td>
<td>87.1±1.0</td>
<td>91.3±0.4</td>
<td>93.1±0.3</td>
</tr>
<tr>
<td>SubG-Con</td>
<td>82.5±0.3</td>
<td>70.8±0.3</td>
<td>73.1±0.5</td>
<td>OOM</td>
<td>OOM</td>
<td>OOM</td>
<td>OOM</td>
</tr>
<tr>
<td><strong>DIMP</strong></td>
<td><strong>83.3±0.5</strong></td>
<td><strong>73.3±0.5</strong></td>
<td><strong>81.4±0.5</strong></td>
<td><strong>83.3±0.4</strong></td>
<td><strong>88.7±0.2</strong></td>
<td><strong>92.1±0.5</strong></td>
<td><strong>94.2±0.4</strong></td>
</tr>
</tbody>
</table>

### Table 4: Node Clustering Results in Terms of NMI and ARI.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Cora NMI</th>
<th>Cora ARI</th>
<th>CiteSeer NMI</th>
<th>CiteSeer ARI</th>
<th>Pubmed NMI</th>
<th>Pubmed ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>0.321</td>
<td>0.230</td>
<td>0.305</td>
<td>0.279</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>Spectral</td>
<td>0.127</td>
<td>0.031</td>
<td>0.056</td>
<td>0.010</td>
<td>0.042</td>
<td>0.002</td>
</tr>
<tr>
<td>BigClam</td>
<td>0.007</td>
<td>0.001</td>
<td>0.036</td>
<td>0.007</td>
<td>0.006</td>
<td>0.003</td>
</tr>
<tr>
<td>GraphEnc</td>
<td>0.109</td>
<td>0.006</td>
<td>0.033</td>
<td>0.010</td>
<td>0.209</td>
<td>0.184</td>
</tr>
<tr>
<td>DeepWalk</td>
<td>0.327</td>
<td>0.243</td>
<td>0.088</td>
<td>0.092</td>
<td>0.279</td>
<td>0.299</td>
</tr>
<tr>
<td>GAE</td>
<td>0.429</td>
<td>0.347</td>
<td>0.176</td>
<td>0.124</td>
<td>0.277</td>
<td>0.279</td>
</tr>
<tr>
<td>VGAE</td>
<td>0.436</td>
<td>0.346</td>
<td>0.156</td>
<td>0.093</td>
<td>0.229</td>
<td>0.213</td>
</tr>
<tr>
<td>MGAE</td>
<td>0.511</td>
<td>0.445</td>
<td>0.412</td>
<td>0.414</td>
<td>0.282</td>
<td>0.248</td>
</tr>
<tr>
<td>ARGA</td>
<td>0.449</td>
<td>0.352</td>
<td>0.350</td>
<td>0.341</td>
<td>0.276</td>
<td>0.291</td>
</tr>
<tr>
<td>ARVGA</td>
<td>0.450</td>
<td>0.374</td>
<td>0.261</td>
<td>0.245</td>
<td>0.117</td>
<td>0.078</td>
</tr>
<tr>
<td>GALA</td>
<td>0.577</td>
<td>0.511</td>
<td>0.441</td>
<td>0.446</td>
<td>0.327</td>
<td>0.321</td>
</tr>
<tr>
<td>MVGRL</td>
<td>0.572</td>
<td>0.495</td>
<td>0.469</td>
<td>0.449</td>
<td>0.322</td>
<td>0.296</td>
</tr>
<tr>
<td><strong>DIMP</strong></td>
<td><strong>0.581</strong></td>
<td><strong>0.522</strong></td>
<td><strong>0.471</strong></td>
<td><strong>0.471</strong></td>
<td><strong>0.346</strong></td>
<td><strong>0.328</strong></td>
</tr>
</tbody>
</table>

### Table 5: Graph Classification Results in Terms of Accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>MUTAG Accuracy</th>
<th>PTC-MR Accuracy</th>
<th>IMDb-B Accuracy</th>
<th>IMDb-M Accuracy</th>
<th>RDT-B Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GraphSage</td>
<td>85.1±7.6</td>
<td>63.9±7.7</td>
<td>72.3±5.3</td>
<td>50.9±2.2</td>
<td>-</td>
</tr>
<tr>
<td>GCN</td>
<td>85.6±5.8</td>
<td>64.2±4.3</td>
<td>74.0±3.4</td>
<td>51.9±3.8</td>
<td>50.0±0.0</td>
</tr>
<tr>
<td>GIN</td>
<td><strong>89.4±5.6</strong></td>
<td><strong>64.6±7.0</strong></td>
<td><strong>75.1±5.1</strong></td>
<td><strong>52.3±2.8</strong></td>
<td><strong>92.1±2.5</strong></td>
</tr>
<tr>
<td>GAT</td>
<td>89.4±6.1</td>
<td><strong>66.7±5.1</strong></td>
<td>70.5±2.3</td>
<td>47.8±3.1</td>
<td>85.2±3.3</td>
</tr>
<tr>
<td>DeepWalk</td>
<td>83.7±1.5</td>
<td>57.9±1.3</td>
<td>50.7±0.3</td>
<td>34.7±0.2</td>
<td>-</td>
</tr>
<tr>
<td>node2vec</td>
<td>72.6±10.0</td>
<td>58.6±8.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sub2vec</td>
<td>61.1±15.0</td>
<td>60.0±6.4</td>
<td>55.3±1.5</td>
<td>36.7±0.8</td>
<td>71.5±0.4</td>
</tr>
<tr>
<td>graph2vec</td>
<td>83.2±9.6</td>
<td>60.2±6.9</td>
<td>71.1±0.5</td>
<td>50.4±0.9</td>
<td>75.8±1.0</td>
</tr>
<tr>
<td>Infograph</td>
<td>89.0±11.1</td>
<td>61.7±14.4</td>
<td>73.0±0.9</td>
<td>49.7±0.5</td>
<td>82.5±1.4</td>
</tr>
<tr>
<td>MVGRL</td>
<td>89.7±1.1</td>
<td>62.5±1.7</td>
<td>74.2±0.7</td>
<td>51.2±0.5</td>
<td>84.5±0.6</td>
</tr>
<tr>
<td>GraphCL</td>
<td>86.8±1.3</td>
<td>71.1±0.4</td>
<td>-</td>
<td>89.5±0.8</td>
<td>-</td>
</tr>
<tr>
<td><strong>DIMP</strong></td>
<td><strong>91.5±1.1</strong></td>
<td><strong>64.2±1.2</strong></td>
<td><strong>74.8±0.8</strong></td>
<td><strong>52.0±0.6</strong></td>
<td><strong>91.9±0.6</strong></td>
</tr>
</tbody>
</table>
Theoretical Analysis

Diverse Message Passing can prevent over-smoothing issue

Semi-supervised Task
\[ h^k_v = \sigma \left( \left( c^k_{vv} \odot h^{k-1}_v \right) + \sum_{u \in \mathcal{N}(v)} c^k_{uv} \odot h^{k-1}_u \right) W^k \].

The connection between learned propagation weights and graph partition

Self-supervised Task
\[ h^k_v = \sigma \left( \left( m^k_{vv} + \sum_{u \in \mathcal{N}(v)} m^k_{uv} \right) W^k \right) \]
\[ m^k_{uv} = \frac{h^{k-1}_v \odot h^{k-1}_u}{\langle h^{k-1}_v, h^{k-1}_u \rangle} \].

The connection between inner-product message and community detection
Theoretical Analysis

Diverse Message Passing can prevent over-smoothing issue

Semi-supervised Task
\[ h_v^k = \sigma \left( \left( \mathbf{c}_{vv} \odot h_v^{k-1} + \sum_{u \in \mathcal{N}(v)} \mathbf{c}_{uv} \odot h_u^{k-1} \right) \mathbf{W}^k \right) \]

The connection between learned propagation weights and graph partition

Self-supervised Task
\[ h_v^k = \sigma \left( \left( \mathbf{m}_{vv}^k + \sum_{u \in \mathcal{N}(v)} \mathbf{m}_{uv}^k \right) \mathbf{W}^k \right) \quad \mathbf{m}_{uv}^k = \frac{h_v^{k-1} \odot h_u^{k-1}}{\langle h_v^{k-1}, h_u^{k-1} \rangle} \]

The connection between inner-product message and community detection
Learned propagation weights vs. Graph partition

**Theorem 1.** The Uniform Message Passing in Eq. (2) with learnable weights $c_{uv}$ is the gradient descent algorithm of the following objective function with node attribute $X$ being the initialization of $H$.

$$
\min_{C, H} \sum_{u,v} \left( b_{uv} c_{uv} + \gamma c_{uv}^2 \right) + 2tr(H^T L_C H),
$$  

where $b_{uv} = g(a_{uv}, \text{dis}(x_i, x_j))$ denotes the similarity between nodes $u$ and $v$, according to both the topology $a_{uv}$ and the distance between attributes $\text{dis}(x_i, x_j)$. $A = [a_{uv}]$ is the adjacency matrix of $G$. $C$ represents the collection of $c_{uv}$, i.e., the adjacency matrix of the learned graph. $L_C$ stands for the Laplacian matrix of the adjacency matrix $C$.

**Theorem 2.** [Ky Fan’s Theorem [30]] There exists

$$
\min_{H \in \mathbb{R}^{N \times F}, H^T H = I} \sum_{f=1}^{F} \sigma_f(L_C),
$$

where $\sigma_f(L_C)$ denotes the $f^{th}$ smallest eigenvalue of the Laplacian matrix $L_C$.

**Theorem 3.** [[31, 32]] The multiplicity $F$ of the eigenvalue $0$ of the Laplacian matrix $L_C$ equals to the number of connected components in the graph, whose similarity matrix is $C$. 

Learned propagation weights vs. Graph partition

**Theorem 1.** The Uniform Message Passing in Eq. (2) with learnable weights $c_{uv}$ is the gradient descent algorithm of the following objective function with node attribute $X$ being the initialization of $H$.

$$h_v^k = \sigma \left( c_{uv}^k h_{uv}^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k h_u^{k-1} \right) W^k,$$

$$\min_{C,H} \sum_{u,v} \left( b_{uv} c_{uv} + \gamma c_{uv}^2 \right) + 2tr(H^T L_C H),$$

s.t. $\forall u \sum c_{uv} = 1$, $0 \leq c_{uv} \leq 1$, $H \in \mathbb{R}^{N \times F}$,

where $b_{uv} = g(a_{uv}, \text{dis}(x_i, x_j))$ denotes the similarity between nodes $u$ and $v$, according to both the topology $a_{uv}$ and the distance between attributes $\text{dis}(x_i, x_j)$. $A = [a_{uv}]$ is the adjacency matrix of $G$. $C$ represents the collection of $c_{uv}$, i.e., the adjacency matrix of the learned graph. $L_C$ stands for the Laplacian matrix of the adjacency matrix $C$.

**Theorem 2.** [Ky Fan’s Theorem [30]] There exists

$$\min_{H \in \mathbb{R}^{N \times F}, H^T H = I} tr(H^T L_C H) = \sum_{f=1}^F \sigma_f(L_C),$$

where $\sigma_f(L_C)$ denotes the $f^{th}$ smallest eigenvalue of the Laplacian matrix $L_C$.

**Theorem 3.** [[31], [32]] The multiplicity $F$ of the eigenvalue 0 of the Laplacian matrix $L_C$ equals to the number of connected components in the graph, whose similarity matrix is $C$. 
Learned propagation weights vs. Graph partition

**Theorem 1.** The Uniform Message Passing in Eq. (2) with learnable weights $c_{uv}$ is the gradient descent algorithm of the following objective function with node attribute $X$ being the initialization of $H$.

$$
\begin{equation}
    h^k_v = \sigma \left( \left( c^k_{uv} h^{k-1}_v + \sum_{u \in N(v)} c^k_{uv} h^{k-1}_u \right) W^k \right),
\end{equation}
$$

$$
\begin{equation}
    \min_{C, H} \sum_{u,v} \left( b_{uv} c_{uv} + \gamma c^2_{uv} \right) + 2\text{tr}(H^T L_C H),
\end{equation}
$$

\begin{equation}
    \text{s.t. } \forall u \sum c_{uv} = 1, 0 \leq c_{uv} \leq 1, H \in \mathbb{R}^{N \times F},
\end{equation}

where $b_{uv} = g(a_{uv}, \text{dis}(x_i, x_j))$ denotes the similarity between nodes $u$ and $v$, according to both the topology $a_{uv}$ and the distance between attributes $\text{dis}(x_i, x_j)$. $A = [a_{uv}]$ is the adjacency matrix of $G$. $C$ represents the collection of $c_{uv}$, i.e., the adjacency matrix of the learned graph. $L_C$ stands for the Laplacian matrix of the adjacency matrix $C$.

**Theorem 2.** [Ky Fan’s Theorem [30]] There exists

$$
\begin{equation}
    \min_{H \in \mathbb{R}^{N \times F}, H^T H = I} \text{tr}(H^T L_C H) = \sum_{f=1}^{F} \sigma_f(L_C),
\end{equation}
$$

where $\sigma_f(L_C)$ denotes the $f^{th}$ smallest eigenvalue of the Laplacian matrix $L_C$.

**Theorem 3.** [[31, 32]] The multiplicity $k$ of the eigenvalue 0 of the Laplacian matrix $L_C$ equals to the number of connected components in the graph, whose similarity matrix is $C$. 

Learned propagation weights vs. Graph partition

**Theorem 1.** The Uniform Message Passing in Eq. (2) with learnable weights $c_{uv}$ is the gradient descent algorithm of the following objective function with node attribute $X$ being the initialization of $H$.

\[
\begin{align*}
\min_{C, H} & \sum_{u,v} (b_{uv}c_{uv} + \gamma c_{uv}^2) + 2\text{tr}(H^T L_C H), \\
\text{s.t.} & \forall u \sum c_{uv} = 1, \ 0 \leq c_{uv} \leq 1, \ H \in \mathbb{R}^{N \times F},
\end{align*}
\]

where $b_{uv} = g(a_{uv}, dis(x_i, x_j))$ denotes the similarity between nodes $u$ and $v$, according to both the topology $a_{uv}$ and the distance between attributes $dis(x_i, x_j)$. $A = [a_{uv}]$ is the adjacency matrix of $G$. $C$ represents the collection of $c_{uv}$, i.e., the adjacency matrix of the learned graph. $L_C$ stands for the Laplacian matrix of the adjacency matrix $C$.

**Theorem 2.** [Ky Fan’s Theorem [30]] There exists

\[
\min_{H \in \mathbb{R}^{N \times F}, H^T H = I} \text{tr}(H^T L_C H) = \sum_{f=1}^{F} \sigma_f(L_C),
\]

where $\sigma_f(L_C)$ denotes the $f^{th}$ smallest eigenvalue of the Laplacian matrix $L_C$.

**Theorem 3.** [[31, 32]] The multiplicity $k$ of the eigenvalue 0 of the Laplacian matrix $L_C$ equals to the number of connected components in the graph, whose similarity matrix is $C$. 

Learned propagation weights vs. Graph partition

\[
\begin{align*}
\mathbf{h}_v^k &= \sigma \left( c_{uv}^k \mathbf{h}_v^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k \mathbf{h}_u^{k-1} \right) \mathbf{W}^k,
\end{align*}
\]
Learned propagation weights vs. Graph partition

**Theorem 4.** The Uniform Message Passing in Eq. (2) actually partitions graph into $F$ connected components based on the similarity $b_{uv} = g(a_{uv}, \text{dis}(x_i, x_j))$ via

$$\min_{C} \sum_{u,v} (b_{uv}c_{uv} + \gamma c_{uv}^2)$$

(12)

$$s.t. \forall u \sum_c c_{uv} = 1, 0 \leq c_{uv} \leq 1, \text{rank}(L_C) = N - F.$$  

(13)

**Theorem 5.** The Diverse Message Passing in Eq. (4) actually partitions graph into 2 connected components (F groups) based on each similarity $b^{(f)}_{uv} = g(a_{uv}, \text{dis}(x_{ij}, x_{jj}))$ via

$$\min_{C^{(f)}} \sum_{u,v} \left(b^{(f)}_{uv}c^{(f)}_{uv} + \gamma (c^{(f)}_{uv})^2 \right), f = 1, ..., F.$$  

(14)

$$s.t. \forall u \sum_c c^{(f)}_{uv} = 1, 0 \leq c^{(f)}_{uv} \leq 1, \text{rank}(L^{(f)}_C) = N - 2.$$  

(15)
Theoretical Analysis

Diverse Message Passing can prevent over-smoothing issue

Semi-supervised Task

$$h^k_v = \sigma \left( \left[ c^k_{vv} \odot h^{k-1}_v \right] + \sum_{u \in \mathcal{N}(v)} \left[ c^k_{uv} \odot h^{k-1}_u \right] W^k \right).$$

The connection between learned propagation weights and graph partition

Self-supervised Task

$$h^k_v = \sigma \left( \left[ m^k_{vv} \right] + \sum_{u \in \mathcal{N}(v)} \left[ m^k_{uv} \right] W^k \right), \quad m^k_{uv} = \frac{h^{k-1}_v \odot h^{k-1}_u}{\langle h^{k-1}_v, h^{k-1}_u \rangle},$$

The connection between inner-product message and community detection
Inner-product message vs Community detection

**Theorem 1.** The diverse and interactive message passing in Eqs. (7) and (8) is equivalent to the expectation-maximization (EM) algorithm to maximize the likelihood of generating graph from community structure \( \{ \mathbf{h}_u \}_{u=1}^N \) via Poisson distribution in (Ball, Karrer, and Newman 2011) as follows

\[
P \left( \mathcal{G} \mid \{ \mathbf{h}_u \}_{u=1}^N \right) = \prod_{u < v} \frac{(\mathbf{h}_u^T \mathbf{h}_v)^{a_{uv}}}{a_{uv}!} \exp \left( -\mathbf{h}_u^T \mathbf{h}_v \right) \tag{13}
\]

\[
\times \prod_u \frac{(1/2 \mathbf{h}_u^T \mathbf{h}_u)^{a_{uu}/2}}{(a_{uu}/2)!} \exp \left( -\frac{1}{2} \mathbf{h}_u^T \mathbf{h}_u \right).
\]

\[
\log P \left( \mathcal{G} \mid \{ \mathbf{h}_u \} \right) \geq \sum_{u,v,z} \left[ a_{uv} q_{uv}(z) \log \frac{h_{uz} h_{vz}}{q_{uv}(z)} - h_{uz} h_{vz} \right]:
\]

**Message**

\[
m_{uv}^k = \frac{\mathbf{h}_v^{k-1} \otimes \mathbf{h}_u^{k-1}}{\langle \mathbf{h}_v^{k-1}, \mathbf{h}_u^{k-1} \rangle},
\]

**Passing**

\[
h_v^k = \sigma \left( m_{vv}^k + \sum_{u \in \mathcal{N}(v)} m_{uv}^k \mathbf{W}^k \right)
\]

\[
q_{uv}(z) = \frac{h_{uz} h_{vz}}{\sum_z h_{uz} h_{vz}} = \frac{h_{uz} h_{vz}}{\langle \mathbf{h}_u, \mathbf{h}_v \rangle}.
\]

\[
h_{uz} = \frac{\sum_v a_{uv} q_{uv}(z)}{\sum_v h_{vz}}.
\]
Conclusions

Semi-supervised Task

\[ h_u^k = \sigma \left( c_{uu}^k \odot h_u^{k-1} + \sum_{u \in \mathcal{N}(v)} c_{uv}^k \odot h_u^{k-1} \right) W^k \]

Learned propagation weights vs. Graph partition

Self-supervised Task

\[ h_v^k = \sigma \left( m_{vv}^k + \sum_{u \in \mathcal{N}(v)} m_{uv}^k \right) W^k \]

\[ m_{uv}^k = \left< h_v^{k-1}, h_u^{k-1} \right> \]

Inner-product message vs. Community detection

Diverse Message Passing can prevent over-smoothing issue