## Beyond Homophily in GNNs: Current Limitations and Effective Designs

### Section 1：Introuction

1. 同质性（Homophily）：连接的节点通常属于同一类或具有相似的特征

异质性/低同质性（Heterophily）：连接的节点可能具有不同类的标签和特征

许多忽略图结构的模型，甚至比GNN（同质性假设）有更好地效果

1. Contributions

①Current Limitations：揭示了GNN在异质性网络上学习的局限性。

②Key Designs for Heterophily & New Model（H2GCN）

**自我和邻居嵌入的分离、高阶邻域、中间表示的组合**，这些设计可以在异质性下促进从图结构的学习。

③Extensive Empirical Evaluation

### Section 2：Notation and Preliminaries

1. **相关定义**

G=（V，E）：无向、无加权图，节点集V和边集E

N（v）：以v为中心的一般邻域

：不包含自我（节点v）的对应邻域

Ni (v)：节点v的i跳邻域

A∈{ 0 , 1 }n×n：邻接矩阵

X∈Rn×F：节点特征矩阵

yv：节点v的标签

半监督节点分类的目标是学习一个映射l：V→Y

**（2）Graph neural networks**

对于每个节点v∈V，存在一个邻域N(v)，使得yv只依赖于自我特征xv和邻域特征{xu：u∈N(v)}。

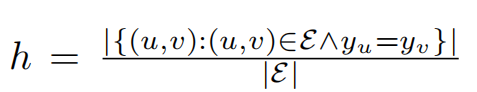
大多数模型通过以下表示学习方法推导出类标签yv：



大多数现有的模型在它们**对邻域N (v)和嵌入函数f的定义**上有所不同。

**（3）Homophily and heterophily**

定义**边的同质比h**作为图同质水平的度量，用它来定义具有强同质性/强异质性的图：

①Definition 1：

h（h的定义**给出了图中所有边的总体趋势**）：是图中连接具有相同类标签（即类内边）的节点的边的分数。

②Definition 2：

具有强同质性的图具有高的边的同质比：h→1；

具有强异质性（即低/弱同质性）的图具有小的边同质比：h→0。

### Section 3：Learning Over Networks with Heterophily 通过具有异质性的网络进行学习

3.1 Effective Designs for Networks with Heterophily 自我和邻居嵌入的分离、高阶邻域、中间表示的组合

**3.1.1 Ego- and Neighbor-embedding Separation 自我和邻居嵌入的分离**

将每个**自我嵌入**（即一个节点的嵌入）**与其邻居的聚合嵌入分开编码**，因为它们在异质性设置中很可能是不同的。形式上，每个节点v在第k轮学习到的表示(或隐藏状态向量)为：



1. **Intuition**

异质性设置中，根据定义Definition2，**一个节点的类标签yv和原始特征xv，及其邻近节点的标签和特征**

**{(yu，xu)：u∈}（直接邻居）可能是不同的**。

然而，典型的GCN设计通过平均或加权平均混合嵌入，作为组合函数，导致任何原始特征的相邻节点（特别是在社区或集群内）的最终嵌入相似。

选择一个COMBINE函数来分离每个节点v及其邻居的表示，允许更多的表达性，其中skipped or non-aggregated representations可以在多轮传播中分别进化，而不会变得非常相似。

1. Theoretical Justification

通过其对测试/训练数据偏差的鲁棒性来衡量其泛化能力。

**3.1.2 Higher-order Neighborhoods 高阶邻域**

在每一轮k中，显式地**聚合每个节点的近邻以外的高阶邻域的信息**：



1. **Intuition**

Definition 3：

①homophily-dominant neighborhoods 同质性主导的社区

如果P( yu = yv | yv ) ≥ P ( yu = y | yv) , ∀u∈N（v）且y ∈ Y≠yv , 则N（v）为预期的同质性主导的社区。

②heterophily-dominant neighborhoods 异质性主导的社区

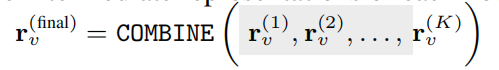
若P( yu = yv | yv ) ＜ P ( yu = y | yv) , 则N（v）为预期的异质性主导的社区。

从这个定义中，预期的同质性主导邻域对GNN层更有利，因为**在这样的邻域中，每个节点v的类标签yv可以由N(v)中的大多数类标签决定。**在异质性的情况下，虽然近邻域可能是异质性主导的，但高阶邻域可能是同质性主导的，可提供更相关的背景。

1. Theoretical Justification

**3.1.3 Combination of Intermediate Representations** **中间表示的组合**

**结合了最后一层上每个节点的中间表示，通过单独利用每个表示的COMBINE函数显式地捕获局部和全局信息。**



1. Intuition

每一轮都在收集不同层的信息，之前的几轮更局部化，而之后的几轮捕获越来越多的全局信息。这种设计更准确地模拟了**低同质性网络中邻域表示的分布**。它还允许类预测利用不同网络中的不同邻域范围，以适应它们的结构属性。

1. Theoretical Justification

**3.2 H2GCN: A Framework for Networks with Homophily or Heterophily**

H2GCN包含三个阶段：**S1：feature embedding，S2：neighborhood aggregation，S3：classification**

1. S1：feature embedding

使用一个图无关的密集层**为每个节点v生成特征嵌入rv(0)**（基于其自我特征xv）：rv(0)=σ(xvWe)，其中σ是一个可选的非线性函数，We是一个可学习的权重矩阵。

1. S2：neighborhood aggregation

①生成的嵌入被聚合，并在节点的邻居内重复更新K轮。



为了**不混合不同的邻域范围**，令COMBINE为连接；AGGR为子邻域 邻居嵌入的度归一化平均值。

②每个节点的最终表示结合了它所有的中间表示（实验表明，串联比最大池化好）：



（3）S3：classification

根据最终嵌入rv(final)对节点进行分类。



Wc是可学习的权重矩阵。

### Section 4：Other Related Work

要注意节点标签的重要性。H2GCN与现有GNN模型的比较：

H2GCN在阶段（S1）中学习一个图无关的特征，并跳过其他模型使用的每轮聚合表示的非线性嵌入（例如，GraphSAGE、MixHop、GCN），从而产生一个更简单而强大的架构。

### Section 5：Empirical Evaluation

Baseline Model：MLP with 1 hidden layer、GCN、GAT、GCN-Cheby、GraphSage、Mixhop

H2GCN：对一阶和二阶邻域（和）进行建模；并考虑两个变体H2GCN-1（K=1），H2GCN-2（K=2）.

**5.1 Evaluation on Synthetic Benchmarks**

（1）Synthetic datasets & setup

生成了具有不同同质性比率h的合成图。

训练、验证和测试分割（25%、25%、50%）

（2）Model comparison

①H2GCN总体上有最好的趋势，在异质性设置中优于大多数基线模型，而在同质性设置中与其他模型持平。

②同质性条件下，GCN、GAT和MixHop，它们混合了自我嵌入和邻居嵌入，性能好。但在异质性下表现差。

③GraphSAGE和GCN-Cheby利用了一些已确定的设计D1-D3，在异质性下，这种设置下更具竞争力。

④异质性下，h∈[0.1,0.3]时，所有模型表现都比较差。

（3）Significance of design choices 通过消融研究证明了D1-D3设计的有效性

（4）The challenging case of low-degree nodes

5.2 Evaluation on Real Benchmarks

（1）Real datasets & setup

（2）Effectiveness of design choices

（3）Additional model comparison

### Section 6：Conclusion