Feature Distribution on Graph Topology Mediates the Effect of Graph Convolution: Homophily Perspective

Soo Yong Lee[1], Sunwoo Kim[1], Fanchen Bu[1], Jaemin Yoo[1], Jiliang Tang[2], Kijung Shin[1]

[1]**KAIST:** {syleetolow, kswoo97, boqvezen97, jaemin, kijungs}@kaist.ac.kr, [2]**MSU:** {tangjili}@msu.edu

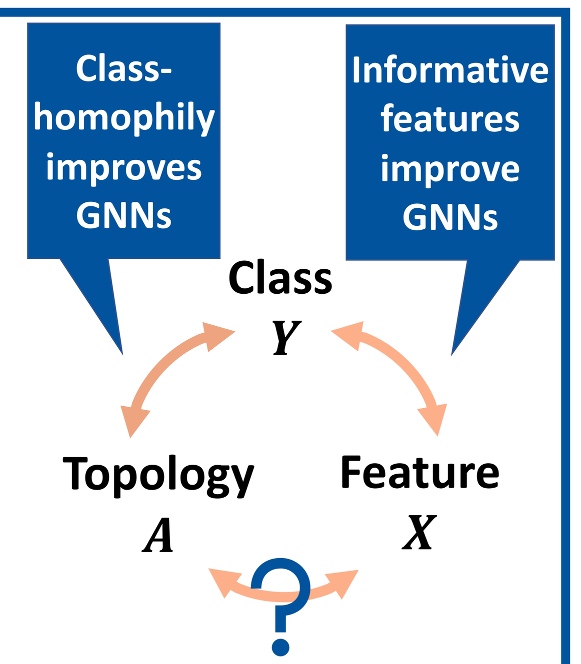
Code and Data: *https://github.com/syleeheal/AX-Dependence-Mediate-Graph-Conv*

Summary

- **Motivation**
  - To understand how dependence between graph topology A and features X (A-X dependence) affect graph convolution
- **Proposed Measure and Random Graph Model**
  - To measure A-X dependence, we propose a measure *class-controlled feature homophily* (CFH)
  - To control CFH with a random graph model, we propose CSBM-X
- **Conclusion**
  - We conclude that A-X dependence (i.e., CFH) mediates the effect of graph conv, such that CFH moderates its force to pull each node feature toward the feature mean of the respective node class, with smaller CFH increasing the force

Introduction

- **Graph Neural Networks (GNNs)**
  - GNNs are functions of graph topology and features
  - Prior studies revealed that GNNs are affected by A-Y relation (e.g., class-homophily) and X-Y relation (e.g., feature informativeness)
  - How the relation between topology A and features X (A-X dependence) affects GNNs is not known



- **Research Question**
  - ***1. How should A-X dependence be measured?***
  - ***2. How does A-X dependence affect GNNs?***

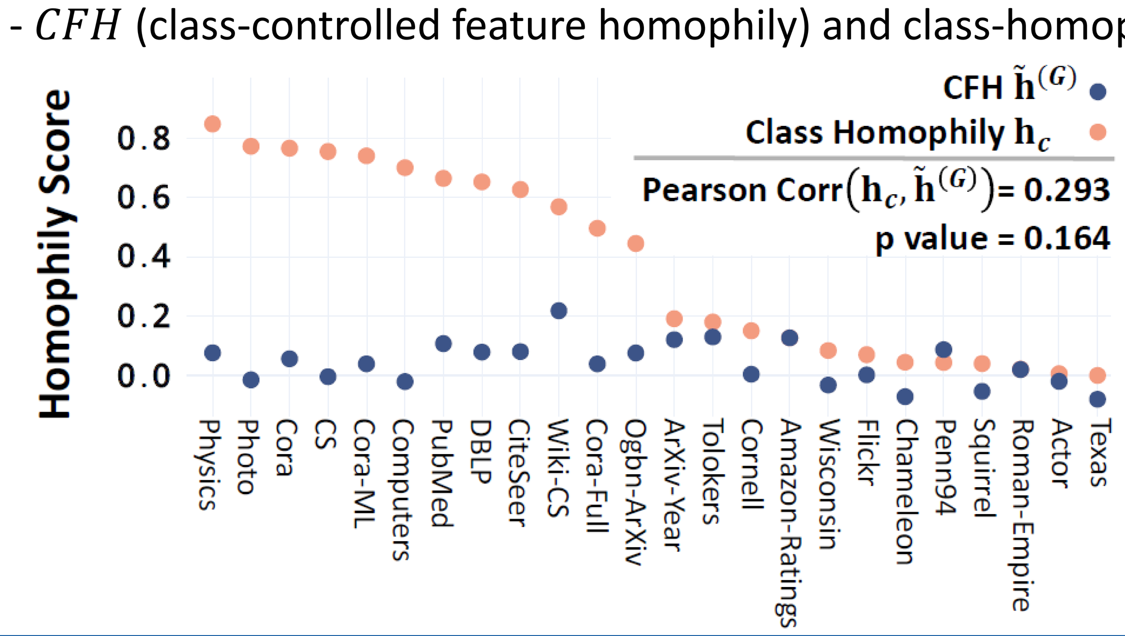
Measure: Class-Controlled Feature Homophily (CFH)

- **Goal**
  - To measure homophily based on node features (i.e., A-X dependence)
- **Handling arbitrary-dimensional, discrete and continuous features**
  - Use *L2 distance* to measure relationship between feature pairs
- **Controlling for a third variable (i.e., node class)**
  - To measure feature homophily while mitigating the effect of a third variable (in our context, node class Y), we define *class-controlled feature* $\boldsymbol{X}_i$

    $\boldsymbol{X}_i = X_i - \left(\frac{1}{|C_i|}\sum_{j\in C_i} X_j\right)$, where C_i is a set of nodes with class Y_i
- **Distinguishing positive, negative, and no dependence**
  - Let $\mathrm{d}_{N(i)}$ be mean distance between $\boldsymbol{X}_i$ and those of neighbors $N(i)$
  - Let $\mathrm{d}_{V\setminus\{v_i\}}$ be mean distance between $\boldsymbol{X}_i$ and those of random nodes $V\setminus\{v_i\}$
  - *Feature homophilic* (positive dependence; $CFH >> 0$), if $\mathrm{d}_{N(i)} \gg \mathrm{d}_{V\setminus\{v_i\}}$
  - *Feature heterophilic* (negative dependence; $CFH << 0$), if $\mathrm{d}_{N(i)} \ll \mathrm{d}_{V\setminus\{v_i\}}$
  - *Feature impartial* (no dependence; $CFH \approx 0$), if $\mathrm{d}_{N(i)} \approx \mathrm{d}_{V\setminus\{v_i\}}$
- **Comparing across different graphs with different features**
  - Class-Controlled Feature Homophily (CFH) = $\begin{cases} 1-\frac{\mathrm{d}_{N(i)}}{\mathrm{d}_{V\setminus\{v_i\}}}, & \text{if } \mathrm{d}_{N(i)} \le \mathrm{d}_{V\setminus\{v_i\}}, \\ \frac{\mathrm{d}_{V\setminus\{v_i\}}}{\mathrm{d}_{N(i)}} - 1, & \text{if } \mathrm{d}_{N(i)} > \mathrm{d}_{V\setminus\{v_i\}}. \end{cases}$
  - Intuitively, a node v_i is $(|CFH|)/(1-|CFH|)$ times closer (or farther) to its neighbors than to random nodes, if $CFH > 0$ (or < 0)
  - For each node v_i, its *distance to random nodes* $\mathrm{d}_{V\setminus\{v_i\}}$ *serves an anchor* to determine CFH magnitude

Observations: Feature Homophily in Real-World Graphs

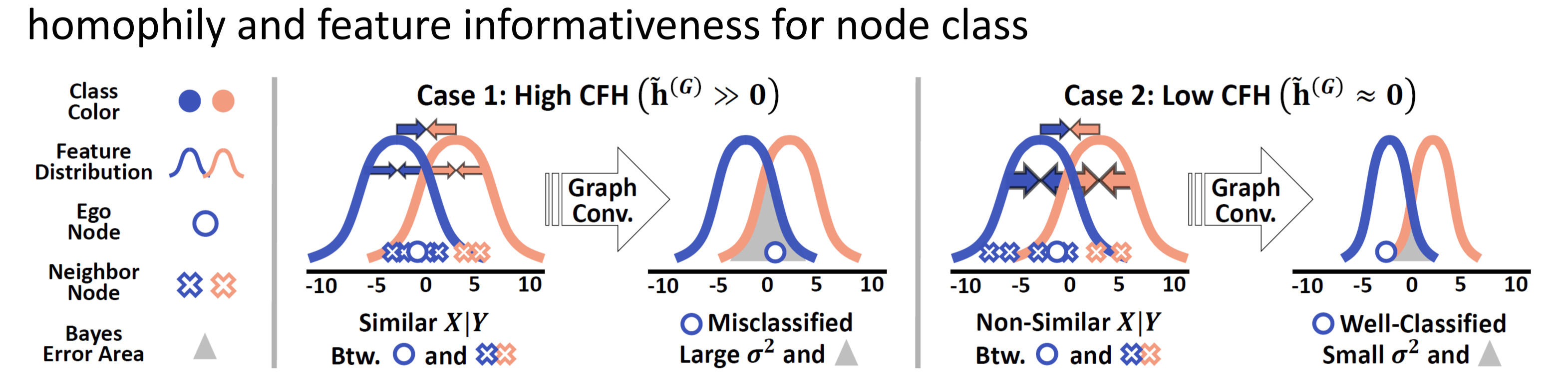
- **Observations**
  - CFH (class-controlled feature homophily) and class-homophily show distinct patterns



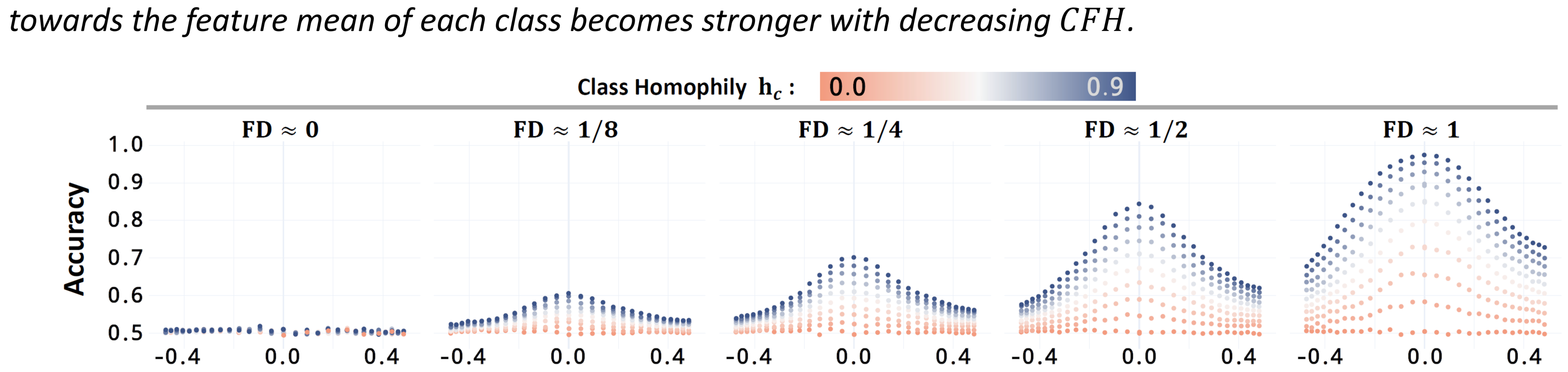
*** Homophily Statistics of Benchmark Graphs.** *(1) The benchmark graphs tend to show small, positive CFH; (2) CFH and class-homophily have a small, positive correlation in the benchmark graphs.*

Theoretical Results in a Random Graph Model CSBM-X

- **Goal**
  - To control CFH (class-controlled feature homophily) with a graph model
- **CSBM-X Overview *(informal)***
  - step 1. sample *nodes* by the model parameter n, i.e., $|V| = n$
  - step 2. divide nodes evenly into *two classes*, i.e, $Y_i \in \{0,1\}$
  - step 3. sample *node features* from Gaussian distributions, i.e, $X_i \sim \mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})$
  - step 4. compute *edges sampling weights* based on node pair features
    - A-X dependence strength parameter τ controls edge sampling weights
    - a positive (or negative) τ exaggerates edge sampling weights among node pairs with higher (or lower) CFH
  - step 5. sample *edges* by the computed weights without replacement, where the model parameters (d^+, d^-) determine same- and different-class neighbor number
- **Key Technical Innovation**
  - CSBM-X can control CFH (i.e., *A-X dependence*), while holding class-homophily (*A-Y dependence*) and feature informativeness (*X-Y dependence*) to be constant
  - By only controlling CFH, CSBM-X provides a suitable theoretical setting to examine the effect of CFH on graph convolution
- **Proposed Theory *(informal)***
  - ***Theorem: A simplified GNN's node classification accuracy increases as A-X dependence strength parameter $\tau \to 0$***
  - Additional result: The effect of τ on GNN performance is moderated by class-homophily and feature informativeness for node class



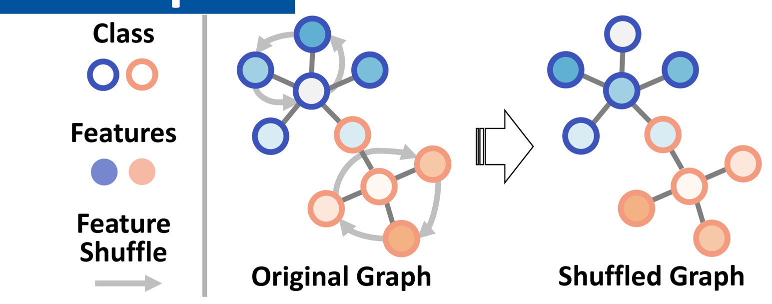
*** A visual intuition of the Theorem.** *When CFH is low (Case 2), the feature distribution of each class shrinks faster (denoted by the arrows) by graph conv., resulting in lower Bayes error. Namely, the power to pull node features towards the feature mean of each class becomes stronger with decreasing CFH.*



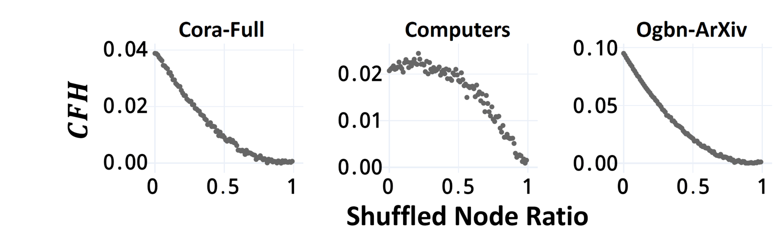
*** A Simplified GNN Performance in CSBM-X Graphs.** *Consistent with the Theorem, the simplified GNN performance increases as $CFH \to 0$ (i.e., $\tau \to 0$). FD roughly indicate feature informativeness for node class.*

Empirical Results in Real-World Graphs

- **Experimental Setting**
  - We shuffle the node features X from the same class Y to reduce CFH (class-controlled feature homophily) of benchmark graphs
  - After the feature shuffle, CFH tends to approach 0 (i.e., no dependence between topology A and features X)
  - The feature shuffle does not interrupt with class-homophily (i.e., *A-Y dependence*) or feature informativeness (i.e., *X-Y dependence*)

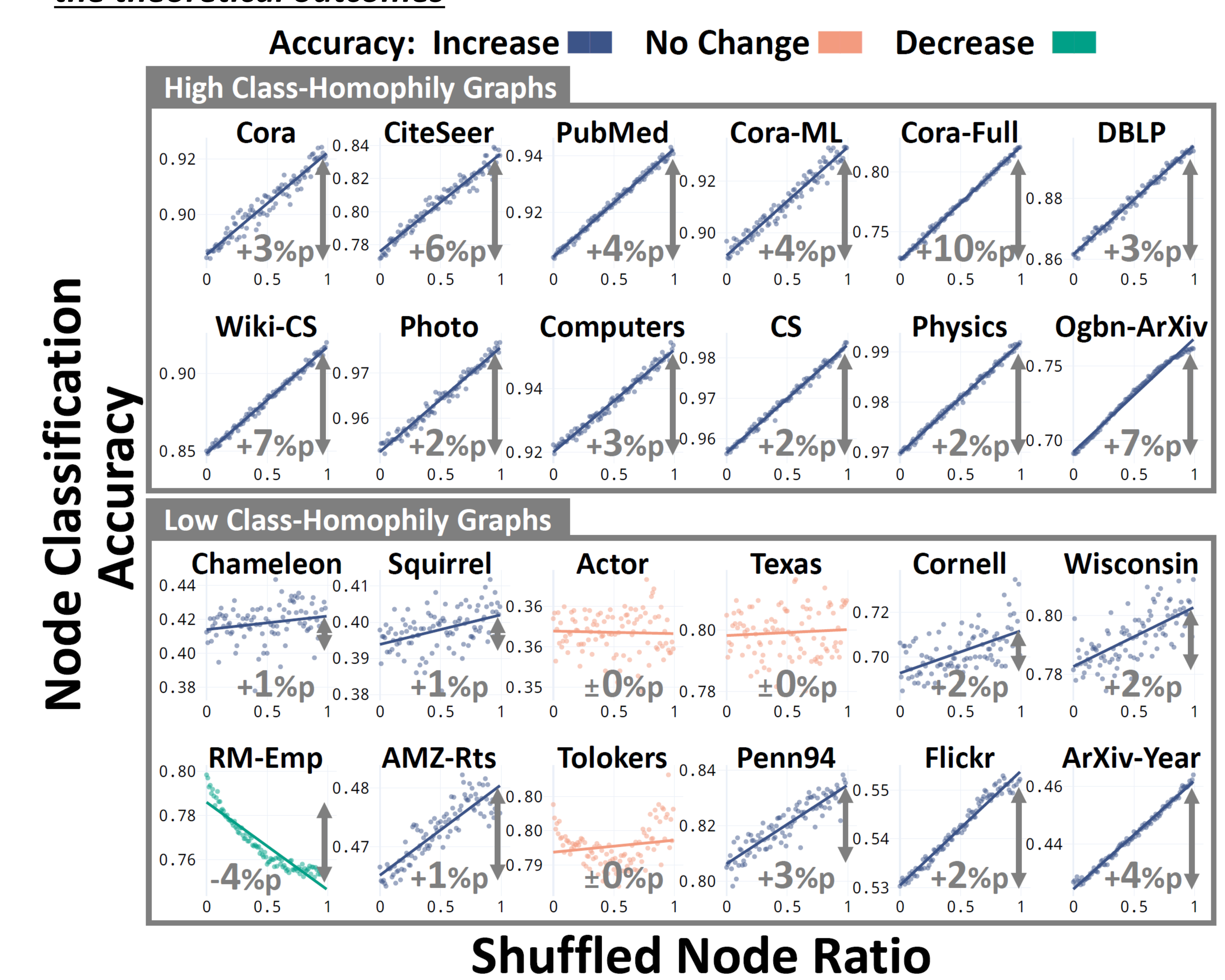


*** Visualization of the Feature Shuffle.** *Features of the same class nodes are shuffled. The shuffled node ratio is 0.6 in the example.*



***** ***CFH after the Feature Shuffle.*** *The feature shuffle reduces CFH to approach 0.*

- **GNN Node Classification After Feature Shuffle**
  - ***The empirical results with 24 real-world graphs are generally consistent with the theoretical outcomes***



*** The Effect of the Feature Shuffle on GNN Node Classification.** (1) The feature shuffle lowers CFH to increase GNN performance; (2) the beneficial effect of lowering CFH is moderated by class-homophily, such that the mean performance increases are 4.4%p vs. 0.5%p.

Discussion

- **Conclusion**
  - We argue that CFH (i.e., class-controlled feature homophily) mediates the effect of graph conv. by moderating the force to pull each node feature toward the feature mean of the respective node class
  - In hindsight, our findings in concert suggest that the recent success of GNNs may have relied on the generally small CFH of the benchmark datasets
  - Looking forward, studying the role of CFH on GNNs is a promising direction
- **Future Directions**
  - *GNN Architecture*: Devising new (1) benchmark datasets with large CFH and (2) GNN models that effectively work on them would be a promising next step
  - *GNN Theory*: The previous GNN theories neglected the role of A-X dependence. We found its impact significant. GNN theories that account for more complex relation between topology and features would enhance understanding of GNNs.