Contrastive Learning as Optimal Homophilic Graph Structure Learning

Pavel Prochazka[1] (✉), Michal Mares[1][0009−0003−7002−274X], and Lukas Bajer[1][0000−0002−9402−6417]
{paprocha,mimares,lubajer}@cisco.com

Cisco, Inc.

Abstract. Contrastive learning has emerged as a dominant paradigm for representation learning, while node homophily–the tendency for similar nodes to be connected–represents a fundamental principle in graph-based learning. This paper establishes a formal connection between these two concepts, revealing that contrastive learning directly optimizes homophilic graph structures through representation learning. Although intuitively related, these frameworks operate in fundamentally different domains: contrastive learning optimizes continuous vector representations, while node homophily is defined on discrete graph structures. To bridge this gap, we introduce a probabilistic framework based on influence matrices that translates between discrete graph adjacencies and continuous similarity relationships. This framework enables us to prove analytically that minimizing contrastive loss is mathematically equivalent to maximizing graph homophily, specifically: $\mathcal{L}^{(i)}_{\text{contrastive}} = -\log L_i$. Our theoretical analysis provides several key insights: (a) standard generalization practices naturally ensure that learned homophily reflects meaningful rather than spurious similarities, (b) an imperfectness parameter ε characterizes how real-world contrastive learning captures both explicit relationships and latent homophily potential, and (c) node selection strategies provide flexibility in applying the framework while maintaining theoretical guarantees. We address current limitations and outline promising directions for future empirical validation across diverse domains. This work establishes rigorous theoretical foundations connecting contrastive learning and graph homophily optimization, opening new directions for adaptive graph construction in graph neural networks.

Keywords: Graph Structure Learning, Homophily, Contrastive Learning, Graph Neural Networks, Representation Learning

1 Introduction

Graph structure learning aims to discover optimal connectivity patterns that capture semantic relationships in data, with homophily–the tendency for similar nodes to be connected–being a fundamental organizing principle. Contrastive learning has demonstrated remarkable effectiveness in representation learning across domains from computer vision to natural language processing, establishing itself as a dominant paradigm for learning meaningful representations through similarity optimization.

This paper establishes the formal connection between these two fundamental concepts, revealing that contrastive learning directly optimizes homophilic graph structures through representation learning. While both frameworks intuitively relate to the principle that similar entities should be grouped together, they operate in fundamentally different domains: contrastive learning optimizes continuous vector representations, while node homophily is defined on discrete graph structures with binary adjacency relationships.

To bridge this gap, we introduce a probabilistic framework based on influence matrices that translates between discrete graph adjacencies and continuous similarity relationships. This framework enables direct comparison between node homophily and contrastive objectives, revealing a fundamental mathematical equivalence that was previously hidden.

Our main theoretical contribution establishes that minimizing per-node contrastive loss is mathematically equivalent to maximizing per-node homophily likelihood when similarities are computed using softmax-normalized dot products. Specifically, we prove that $\mathcal{L}^{(i)}_{\text{contrastive}} = -\log L_i$, demonstrating that contrastive learning has been implicitly optimizing graph homophily all along.

Key theoretical results. We present two fundamental theorems: First, our probabilistic homophily likelihood formulation generalizes traditional node homophily, reducing to the classical definition when influence weights correspond to uniform adjacency relationships. Second, contrastive loss equals the negative log-likelihood of this homophily measure, establishing the direct mathematical connection between representation learning and graph structure optimization.

Framework insights and practical considerations. Our analysis provides several important insights: the equivalence explains why contrastive learning is effective for similarity-based tasks from a graph-theoretic perspective; standard generalization practices naturally ensure that learned homophily reflects meaningful rather than spurious similarities; and the ε-imperfectness parameter reveals how representation learning captures both explicit relationships and latent homophily potential. The influence matrix construction naturally connects to weighted k-nearest neighbor classification, demonstrating practical applications of the learned homophilic structures.

Scope and future directions. While our theoretical analysis establishes the fundamental mathematical connection, our current scope focuses on the core theoretical framework with softmax-normalized similarities. Several practical aspects warrant future empirical investigation, including the impact of temperature parameters, batch sampling effects, and regularization mechanisms on learned homophilic structures. Additionally, domain-specific applications to graph neural networks and adaptive graph construction methods represent promising directions for translating these theoretical insights into practical improvements. We address these limitations and outline specific future research directions in our comprehensive analysis of practical considerations.

This work provides rigorous theoretical foundations connecting contrastive learning and graph homophily optimization, opening new directions for understanding representation learning through graph-theoretic principles while suggesting applications to adaptive graph construction in graph neural networks and similarity-based learning methods.

2 Related Work

2.1 Graph Structure Learning and Homophily

Graph structure learning aims to discover or optimize connectivity patterns in data, with homophily-the principle that similar nodes tend to be connected-serving as a fundamental organizing concept [9]. Classical approaches include methods for learning graph topology from data [4] and structure learning for graphical models [7]. Recent work has focused on adaptive graph construction for graph neural networks, where the graph structure itself becomes learnable [5].

Homophily has been extensively studied in social networks [9] and has become central to understanding graph neural network performance [18]. Methods for measuring and leveraging homophily include homophily-aware graph neural networks [18] and adaptive approaches that handle both homophilic and heterophilic relationships [8]. Our work contributes to this area by showing how contrastive learning implicitly optimizes homophilic graph structures.

2.2 Contrastive Learning and Graph Representation

Contrastive learning has emerged as a dominant paradigm in representation learning [2], with recent extensions to graph data showing remarkable success. Graph contrastive learning methods like GraphCL [16] and GRACE [20] apply contrastive objectives to learn node and graph representations by contrasting different augmented views of graph structures.

These methods demonstrate that contrastive objectives can effectively capture structural relationships in graphs, though the theoretical connection between contrastive learning and explicit graph structure optimization remains underexplored. Our work provides theoretical foundations for understanding how contrastive learning optimizes homophilic relationships, complementing empirical successes in graph contrastive learning.

2.3 Similarity-Based Classification and Graph Methods

The connection between similarity-based classification and graph methods has been explored in various contexts. Graph-based semi-supervised learning [19] propagates labels through similarity graphs, while spectral clustering methods [12] use graph Laplacians derived from similarity matrices. K-nearest neighbor classification can be viewed as operating on similarity graphs where edges connect each query to its nearest neighbors [3].

Recent work has explored learnable similarity metrics for KNN [14] and probabilistic formulations [6]. However, the connection between these similarity-based approaches and homophilic graph structure learning has not been theoretically established. Our framework bridges this gap by showing how contrastive learning optimizes homophilic structures that support effective similarity-based inference.

2.4 Theoretical Connections in Graph Learning

Several works have explored theoretical foundations of graph-based learning methods. Belkin and Niyogi [1] established connections between graph Laplacians and manifold learning, while Zhou et al. [17] provided theoretical analysis of graph-based semi-supervised learning. More recently, theoretical understanding of graph neural networks has focused on expressivity [15] and generalization [11].

The theoretical connection between contrastive learning and graph structure optimization has received limited attention. Tian et al. [10] analyzed contrastive representations for linear classification, while Wang and Isola [13] studied alignment and uniformity properties. Our work extends these theoretical insights to the graph domain, showing how contrastive objectives implicitly optimize homophilic graph structures.

Our contribution differs from existing work by establishing a theoretical connection between contrastive learning and homophilic graph structure learning, providing a graph-theoretic perspective on why contrastive methods are effective while suggesting applications to adaptive graph construction in graph neural networks

3 Problem Statement and Notation

3.1 Notation

We consider a graph-based view where data is represented as a graph $G = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V}$ corresponding to examples and edges $\mathcal{E}$ capturing relationships between them. The adjacency matrix

$\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ encodes pairwise relationships, where $A_{ij} = 1$ indicates an edge between nodes i and j, and $A_{ij} = 0$ otherwise. The neighborhood of node i is defined as $\mathcal{N}(i) = \{j \in \mathcal{V} : A_{ij} = 1\}$.

From a labeling perspective, each node i is associated with a label $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ represents the set of possible labels. The complete node set is denoted as $\mathcal{V} = \{1, 2, \ldots, n\}$. Additionally, nodes can be characterized by their feature representations, where each node i is described by features $\mathbf{x}_i \in \mathbb{R}^d$. Through a learnable mapping $f : \mathbb{R}^d \to \mathbb{R}^k$, node features are transformed into k-dimensional representations $\mathbf{z}_i = f(\mathbf{x}_i)$ in the learned embedding space.

3.2 Core Definitions

Two fundamental concepts form the basis of our analysis. First, we define node homophily as the fraction of neighbors sharing the same label as the focal node:

$$H_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \delta(y_i = y_j) \tag{1}$$

where $\delta(\cdot)$ is the indicator function. This measure quantifies the local homophily around each node, with higher values indicating stronger label consistency within the neighborhood.

Second, we consider the contrastive loss for a node i, which optimizes representations by encouraging similarity between same-class examples while promoting dissimilarity between different-class examples:

$$\mathcal{L}^{(i)}_{\text{contrastive}} = -\log \left(\frac{\sum_j \delta(y_i = y_j) \exp(\mathbf{z}_i^T \mathbf{z}_j / \tau)}{\sum_j \exp(\mathbf{z}_i^T \mathbf{z}_j / \tau)} \right) \tag{2}$$

where τ is the temperature parameter controlling the concentration of the similarity distribution. The summations are computed over a training batch containing both positive pairs (same-class examples) and negative pairs (different-class examples) [2].

3.3 Problem Statement

The central challenge we address lies in understanding the relationship between these two seemingly distinct formulations. Node homophily operates on discrete graph structures with binary adjacency relationships, measuring label consistency within fixed neighborhoods. In contrast, contrastive loss operates on continuous representation spaces, optimizing learned embeddings through similarity-based objectives without explicit reference to graph topology. This fundamental disconnect creates a theoretical gap: while both concepts intuitively relate to the principle that similar entities should be grouped together, their mathematical formulations operate in different domains–discrete graphs versus continuous representations (see Figure 1). The key obstacle is that homophily requires a materialized graph structure with defined edges, while contrastive learning produces similarity relationships in embedding space without explicit graph construction.

Our research question addresses this gap directly: *Can we establish a formal mathematical relationship between contrastive loss optimization and homophilic graph structure learning?* Specifically, we seek to understand whether contrastive learning implicitly optimizes for homophilic relationships, and if so, under what conditions this equivalence holds.

Resolving this question has significant theoretical and practical implications. Theoretically, it would provide a graph-theoretic foundation for understanding why contrastive learning is effective

for similarity-based tasks. Practically, it could inform the design of graph neural networks and adaptive graph construction methods by revealing how representation learning naturally encodes structural relationships.

To bridge this gap, we propose a probabilistic framework that translates between discrete graph structures and continuous similarity relationships, enabling direct comparison between node homophily and contrastive objectives. This framework reveals that contrastive learning can be viewed as optimizing homophilic graph structures through representation learning, with practical applications to weighted similarity-based classification methods.

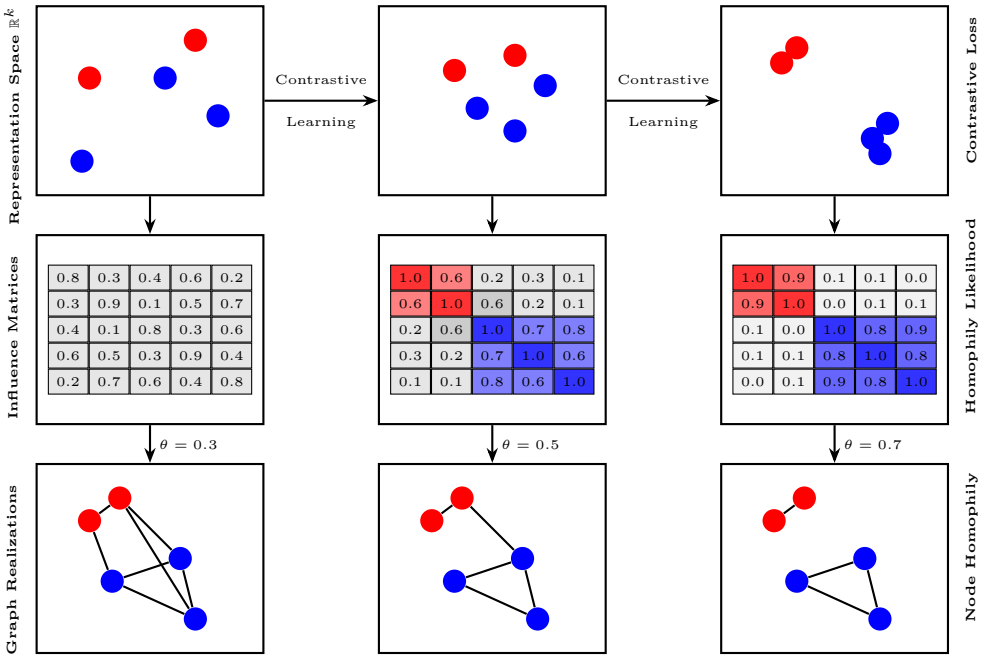

Fig. 1. Illustration of the equivalence between contrastive learning and homophilic graph optimization. The top row shows the natural progression of contrastive learning in representation space $\mathbb{R}^k$, where nodes (red and blue circles representing different classes) evolve from random distribution to well-separated clusters. The middle and bottom rows demonstrate the consequences of this representation learning progression. As contrastive learning separates node representations, the influence matrices (middle row) exhibit increasingly homophilic patterns, with stronger within-class similarities (darker red/blue cells) and weaker cross-class similarities (lighter gray cells). The bottom row shows corresponding graph realizations under different thresholds θ, demonstrating how better-separated representations lead to improved node homophily in the materialized graphs. The left labels identify the three operational spaces: representation space, influence matrices, and graph realizations. The right labels indicate the corresponding measures: contrastive loss $\mathcal{L}^{(i)}_{\text{contrastive}}$, homophily likelihood L_i, and node homophily H_i. Our theoretical results (Theorems 1 and 2) establish the mathematical equivalences between these measures, showing that minimizing contrastive loss is equivalent to maximizing homophily likelihood, which generalizes traditional node homophily.

4 Bridging Representations and Graph Structures

To address the fundamental challenge identified in the previous section, we develop a probabilistic framework that enables direct comparison between discrete graph homophily and continuous representation-based contrastive learning. This framework centers on the concept of an influence matrix that captures pairwise relationships in a continuous manner, providing the necessary bridge between these two domains.

4.1 Motivation for Probabilistic Formulation

The core difficulty in connecting node homophily and contrastive loss lies in their operational domains. While homophily is defined on discrete graphs with binary adjacency relationships, contrastive loss operates on continuous representation spaces through similarity-based objectives. To bridge this gap, we require a unified framework that can express both discrete structural relationships and continuous similarity patterns.

Our proposed solution introduces a probabilistic formulation based on an influence matrix– essentially a weighted adjacency matrix that can capture both discrete and continuous relationships. This influence matrix serves three critical functions, as illustrated in Figure 1: it links discrete graph structures to continuous representations through distances or similarities, provides an intuitive foundation for a continuous version of homophily, and aligns with traditional node homophily under specific conditions that we will demonstrate theoretically.

The influence matrix approach allows us to generalize beyond binary edge relationships to weighted connections that reflect the strength of similarity or influence between nodes. This generalization is essential for understanding how contrastive learning, which naturally produces continuous similarity measures, relates to graph-theoretic concepts like homophily.

4.2 Influence Matrix Construction

The influence matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ captures pairwise relationships between nodes, where each entry W_{ij} represents the influence of node j on node i. This formulation is flexible enough to accommodate various similarity measures derived from learned representations $\mathbf{z}_i$ and $\mathbf{z}_j$.

The influence weights can be constructed using different similarity functions. For Euclidean distance-based measures, we define $W_{ij} = g(\|\mathbf{z}_i - \mathbf{z}_j\|)$ where $g(\cdot)$ is a decreasing function that converts distances to influence weights. Alternatively, for dot product-based similarities, we use $W_{ij} = g(\mathbf{z}_i^T \mathbf{z}_j)$ where $g(\cdot)$ is an increasing function. The specific choice of $g(\cdot)$ depends on the application, but common choices include exponential functions that naturally produce probability-like distributions.

For analysis purposes, we focus on a single row of the influence matrix, representing the weight vector $\mathbf{w}_i = (w_1, w_2, \ldots, w_M)$ where $w_j = W_{ij}$ represents the influence of node j on node i. For convinience, we assume $W_{ii} = 0$ for all i, meaning a node does not influence itself, which is a common assumption in many graph-based learning scenarios.

4.3 Per-Node Homophily Likelihood

Building on the influence matrix framework, we introduce the concept of per-node homophily likelihood, which generalizes traditional node homophily to continuous influence weights. For a node i with label y_i, the homophily likelihood is defined as:

$$L(y_i = c \mid \mathbf{y}, \mathbf{W}) = \sum_j w_j \cdot \delta(y_j = c), \tag{3}$$

where $\delta(y_j = c)$ is the indicator function that equals 1 if $y_j = c$ and 0 otherwise. This formulation captures the total influence weight that nodes with label c exert on node i.

The proper normalization of the influence matrix is crucial for the probabilistic interpretation of homophily likelihood. When the influence matrix is row-normalized such that $\sum_j w_j = 1$ for all i, the homophily likelihood naturally expresses the probability that node i has label c based on weighted voting from all other nodes. Under this normalization, the likelihood represents how much influence weight mass concentrates on same-class neighbors. If we consider the weights to same-class nodes as $w_1, w_2, \ldots, w_\ell$ and weights to different-class nodes as $w_{\ell+1}, w_{\ell+2}, \ldots, w_m$, then the homophily likelihood becomes

$$\frac{w_1 + \cdots + w_\ell}{w_1 + \cdots + w_m} = w_1 + \cdots + w_\ell$$

(since the denominator equals 1 under normalization). This mass concentration interpretation provides an intuitive understanding: higher concentration of weight mass on same-class neighbors corresponds to higher homophily and higher likelihood values.

The homophily likelihood offers several advantages over traditional binary homophily measures. As shown in Figure 1,it naturally handles continuous influence weights, enabling analysis of representation-based similarities. It aligns with node-level analysis, making it compatible with per-node homophily definitions. Most importantly, it provides the weight concentration interpretation described above, where the likelihood directly quantifies the proportion of influence mass concentrated on same-class neighbors.

This probabilistic formulation has a natural connection to similarity-based classification methods. The resulting classification rule $\hat{y}_i = \arg\max_c \sum_j w_j \cdot \delta(y_j = c)$ corresponds directly to weighted k-nearest neighbor classification. This connection demonstrates how learned homophilic structures naturally support similarity-based inference, with the influence matrix $\mathbf{W}$ parameterizing the weighting scheme derived from the optimized graph structure.

4.4 Theoretical Results

With the probabilistic framework established, we can now present our main theoretical contributions that formally connect homophily likelihood to both traditional node homophily and contrastive learning objectives.

Theorem 1 (Homophily-Likelihood Equivalence). *For an influence matrix $\mathbf{W}$ derived from graph adjacency with row-normalized weights $W_{ij} = 1/|\mathcal{N}(i)|$ for $j \in \mathcal{N}(i)$ and $W_{ij} = 0$ otherwise, the homophily likelihood defined in equation (3) equals the node homophily defined in equation (1).*

Proof: Given the influence matrix $\mathbf{W}$ with $W_{ij} = 1/|\mathcal{N}(i)|$ for $j \in \mathcal{N}(i)$ and $W_{ij} = 0$ otherwise, substituting into Equation (3) yields Equation (1). □

This theorem establishes that our probabilistic formulation is a proper generalization of traditional node homophily. When the influence matrix corresponds to uniform weights over graph neighbors, the likelihood measure reduces exactly to the standard homophily definition.

Theorem 2 (Contrastive-Homophily Equivalence). *For softmax-normalized similarities*

$$W_{ij} = \frac{\exp(\mathbf{z}_i^T \mathbf{z}_j/\tau)}{\sum_k \exp(\mathbf{z}_i^T \mathbf{z}_k/\tau)},$$

the per-node contrastive loss equals the negative log-likelihood of homophily: $\mathcal{L}^{(i)}_{contrastive} = -\log L_i$.

Proof: Given the softmax-normalized weights $W_{ij} = \frac{\exp(\mathbf{z}_i^T \mathbf{z}_j/\tau)}{\sum_k \exp(\mathbf{z}_i^T \mathbf{z}_k/\tau)}$, the homophily likelihood becomes $L_i = \sum_{j \in \text{same class}} W_{ij}$. By the definition of contrastive loss in equation (2), we have $\mathcal{L}^{(i)}_{\text{contrastive}} = -\log(L_i)$, establishing the direct equivalence. □

These theorems lead directly to our main result:

Corollary 1 (Contrastive Learning Optimizes Homophily). *Minimizing the per-node contrastive loss is equivalent to maximizing the per-node homophily likelihood. Therefore, contrastive learning directly optimizes homophilic graph structures through representation learning.*

This corollary provides the theoretical foundation for understanding why contrastive learning is effective for tasks that benefit from homophilic relationships, and establishes a formal connection between representation learning and graph structure optimization. Figure 1 visualizes this progression, showing how contrastive learning naturally leads to more homophilic graph structures.

5 Practical Considerations for Homophily Likelihood

Having established the theoretical equivalence between contrastive learning and homophily optimization, we now address practical aspects that arise when applying this framework to real-world scenarios. The following key questions emerge in the context of contrastive learning: when learned homophily is meaningful rather than just grouping of examples with the same label, how do continuous influence weights from learned representations relate to discrete graph structures, and which nodes should be included when computing homophily likelihood in practice?

5.1 Generalization and Meaningful Homophily

While our theoretical equivalence demonstrates that contrastive learning optimizes homophily likelihood, a fundamental question remains: does this optimization produce meaningful homophilic structures? The homophily likelihood L_i depends only on labels and influence weights, without direct reference to input representations. This raises the concern that a model could achieve high homophily likelihood by connecting any same-labeled examples, regardless of their underlying similarity in the input space.

However, no additional mechanisms beyond standard generalization practices are needed to ensure meaningful homophily. When contrastive learning generalizes well to unseen data, the learned representations must capture genuine similarity patterns rather than spurious label correlations. A model that memorizes arbitrary same-label connections without learning meaningful features would fail to maintain high homophily likelihood on new examples, as the influence matrix computed from learned representations would not reflect true similarities for unseen data. Thus, good generalization performance serves as a natural indicator that the learned homophilic structures capture meaningful rather than spurious relationships.

This connection between generalization and homophily quality reinforces that standard regularization and validation practices naturally ensure that the learned homophilic structures reflect meaningful similarities. The theoretical equivalence thus provides both the optimization target (homophily likelihood) and the quality assurance mechanism (generalization) for learning meaningful graph structures through contrastive learning.

5.2 From Discrete Graphs to Continuous Representations

The transition from discrete graph adjacencies to continuous representation-based similarities requires careful analysis. While traditional graph homophily operates with binary edge weights $W_{ij} \in \{0, 1\}$, contrastive learning naturally produces continuous similarity measures $W_{ij} \in \mathbb{R}^+$. To understand this transition, we introduce an imperfectness parameter ε that models how learned representations might deviate from perfect discrete adjacencies.

Consider a scenario where influence weights are defined as $W_{ij} = 1 - \varepsilon$ for connected neighbors $j \in \mathcal{N}(i)$ and $W_{ij} = \varepsilon$ for non-neighbors $j \notin \mathcal{N}(i)$. This parameter $\varepsilon \in [0, 1]$ is not part of the actual method but serves as an analytical tool for understanding the relationship between discrete and continuous formulations. When $\varepsilon \to 0$, we recover the exact correspondence between homophily likelihood and traditional node homophily, while finite ε values represent the realistic continuous similarities produced by representation learning.

To analyze the impact of this parameter, we decompose the likelihood calculation into four cases based on node relationships. For a given node i, we define:

- c^+: same-class neighbors (connected nodes with matching labels)
- c^-: same-class non-neighbors (disconnected nodes with matching labels)
- d^+: different-class neighbors (connected nodes with different labels)
- d^-: different-class non-neighbors (disconnected nodes with different labels)

Under the ε-parameterized influence weights, the homophily likelihood becomes:

$$L_i = \frac{c^+ \cdot (1 - \varepsilon) + c^- \cdot \varepsilon}{c^+ \cdot (1 - \varepsilon) + c^- \cdot \varepsilon + d^+ \cdot (1 - \varepsilon) + d^- \cdot \varepsilon} \tag{4}$$

Comparing this to traditional node homophily $H_i = \frac{c^+}{c^+ + d^+}$, we can verify that $L_i = H_i$ when $\varepsilon \to 0$, confirming our theoretical framework. However, for finite ε values, the likelihood captures both the "understood part" (traditional edge-based homophily) and an "extended part" that reflects latent homophily potential from non-neighbor relationships.

This extended interpretation provides valuable insights into representation learning. When adding a same-class node without an edge to the graph, traditional homophily remains unchanged, but the likelihood slightly increases as it captures the potential for higher homophily through the c^- term. Conversely, adding a different-class non-neighbor slightly decreases the likelihood through the d^- term, reflecting a dilution of homophily potential. The parameter ε controls the balance between these understood and extended components, with edge changes having large impacts on both measures while node additions primarily affecting the likelihood through the extended terms.

This analysis demonstrates how real-world similarities from learned representations naturally relate to our theoretical framework through both understood and extended homophily components. The parameter ε serves as an analytical lens for understanding this relationship–in practice, it emerges from the imperfectness of learned representations rather than being explicitly controlled.

The connection between representation quality and effective ε values shows how the theoretical equivalence captures both explicit relationships and latent homophily potential in learned representations.

5.3 Node Selection for Likelihood Computation

A practical question arises regarding which nodes to include when computing homophily likelihood according to Equation (3). The summation does not explicitly specify the node set, and this choice has important implications for both computational efficiency and theoretical interpretation.

From a practical perspective, the primary use case involves learning representations to construct homophilic graph structures, rather than analyzing existing graphs. When deriving influence matrices from learned representations, the key decision concerns how to construct the final graph for homophily evaluation. This typically involves selecting a threshold on influence weights to identify the most influential nodes, which defines the effective neighborhood for each node in the constructed graph.

The threshold selection creates a trade-off between graph density and connection strength. Higher thresholds produce sparser graphs by retaining only the strongest similarity connections, while lower thresholds yield denser graphs that include weaker similarity relationships. From this perspective, the node selection for likelihood computation should align with the expected neighborhood size in the final constructed graph.

This practical consideration suggests a natural approach: when computing homophily likelihood for evaluation purposes, select the top-k most influential nodes for each focal node, where k corresponds to the desired average degree in the target graph structure. This approach ensures that the likelihood computation reflects the same structural constraints that will be applied in the final graph construction, providing a more meaningful assessment of the learned homophilic relationships.

5.4 Framework Limitations, Scope and Future Directions

Our theoretical framework establishes a fundamental connection between contrastive learning and homophily optimization, but its current scope is defined by several key assumptions and limitations that naturally point toward future research directions.

Methodological Assumptions. Our main theoretical results assume softmax-normalized similarities, which, while widely used in contrastive learning methods like SimCLR and InfoNCE, represents a specific normalization scheme. Similarly, our analysis focuses on standard contrastive loss formulations without explicit treatment of regularization effects, which may influence the learned homophilic structures in practice. These assumptions define the current theoretical scope while highlighting natural extensions: analyzing other similarity measures and normalization approaches, and investigating how different regularization mechanisms affect the contrastive-homophily equivalence.

Node Selection and Homophily Interpretation. While our framework establishes equivalence for any selection of nodes used in likelihood computation, providing considerable flexibility in practical applications, the meaningfulness of the learned homophily depends critically on node selection strategies. The choice of which nodes to include when computing homophily likelihood– whether through batch sampling, k-nearest neighbor selection, or threshold-based filtering–affects the relevance of the computed likelihood to true homophilic relationships in the data. Although the mathematical equivalence $\mathcal{L}^{(i)}_{\text{contrastive}} = -\log L_i$ holds regardless of node selection strategy, the interpretability of the resulting homophily measures varies significantly. This limitation suggests that

node selection strategies should prioritize the most similar examples according to learned representations, though the impact of different selection approaches on homophily quality requires deeper empirical investigation.

Domain-Specific Applications and Empirical Validation. Our current work establishes the theoretical foundation without addressing domain-specific implementations, representing a key limitation in scope. The framework's behavior across different data modalities–images, text, tabular data, and existing graph structures–remains an open question with significant practical implications. In the graph domain specifically, our theoretical insights suggest applications to task-aware graph rewiring, where existing graph structures could be adaptively modified based on learned homophilic relationships. However, understanding how the theoretical equivalence translates to improved performance in specific domains requires comprehensive empirical investigation.

These limitations naturally define future research agenda: extending the theoretical analysis to broader classes of similarity measures and contrastive learning variants, developing principled node selection strategies with theoretical guarantees, and conducting systematic empirical validation across diverse domains and applications. The domain-specific investigations are particularly promising, as they could demonstrate how the theoretical framework enables practical improvements in graph neural networks, similarity-based classification, and adaptive graph construction methods.

6 Conclusion

We have established a rigorous theoretical connection between contrastive learning and homophilic graph structure optimization through a probabilistic framework based on influence matrices. Our main contribution proves that contrastive loss optimization is mathematically equivalent to maximizing homophily likelihood, revealing that contrastive learning directly optimizes homophilic graph structures through representation learning.

Our theoretical framework makes several key contributions. We introduce a probabilistic formulation of homophily that generalizes traditional node homophily to continuous influence weights, enabling analysis of representation-based similarities. Through two fundamental theorems, we prove that our homophily likelihood reduces to classical node homophily under uniform adjacency weights, and that contrastive loss equals the negative log-likelihood of homophily for softmax-normalized similarities. These results establish the mathematical identity $\mathcal{L}^{(i)}_{\text{contrastive}} = -\log L_i$, demonstrating that minimizing contrastive loss is equivalent to maximizing homophily likelihood. The influence matrix framework provides a general foundation extensible to other similarity measures beyond softmax normalization.

The theoretical insights provide important implications for understanding representation learning. The equivalence explains why contrastive learning is effective for similarity-based tasks from a graph-theoretic perspective. The influence matrix construction naturally connects to weighted k-nearest neighbor classification, demonstrating practical applications of learned homophilic structures. Our analysis reveals how representation learning captures both explicit relationships and latent homophily potential, with standard generalization practices naturally ensuring meaningful rather than spurious similarities.

While our theoretical analysis establishes the fundamental mathematical connection, comprehensive empirical validation across diverse domains remains essential for translating these insights into practical improvements. This work provides a rigorous mathematical foundation connecting

contrastive learning and graph homophily optimization, opening new theoretical directions for understanding representation learning through graph-theoretic principles and suggesting promising applications to adaptive graph construction in graph neural networks and similarity-based learning methods.

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation **15**(6), 1373–1396 (2003)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning (2020)
3. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory (1967)
4. Dong, X., Thanou, D., Rabbat, M., Frossard, P.: Learning graphs from data: A signal representation perspective. IEEE Signal Processing Magazine **36**(3), 44–63 (2019)
5. Franceschi, L., Niepert, M., Pontil, M., He, X.: Learning discrete structures for graph neural networks. In: International conference on machine learning. pp. 1972–1982. PMLR (2019)
6. Holmes, C.C., Adams, N.M.: A probabilistic nearest neighbour method for statistical pattern recognition. Journal of the Royal Statistical Society: Series B (2002)
7. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. MIT press (2009)
8. Lim, D., Hohne, F., Li, X., Huang, S.L., Gupta, V., Bhalerao, O., Lim, S.N.: Large scale learning on non-homophilic graphs: New benchmarks and strong simple methods. In: Advances in Neural Information Processing Systems. vol. 34, pp. 20887–20902 (2021)
9. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology **27**(1), 415–444 (2001)
10. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? In: Advances in Neural Information Processing Systems (2020)
11. Verma, S., Zhang, Z.L.: Stability and generalization of graph convolutional neural networks. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1539–1548 (2019)
12. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**(4), 395–416 (2007)
13. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning (2020)
14. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of machine learning research (2009)
15. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations (2019)
16. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y.: Graph contrastive learning with augmentations. In: Advances in Neural Information Processing Systems. vol. 33, pp. 5812–5823 (2020)
17. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency **16** (2004)
18. Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., Koutra, D.: Beyond homophily in graph neural networks: Current limitations and effective designs. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7793–7804 (2020)
19. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions pp. 912–919 (2003)
20. Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., Wang, L.: Deep graph contrastive representation learning. In: ICML Workshop on Graph Representation Learning and Beyond (2020)