OPTIMAL TRANSPORT-GUIDED GEOMETRIC CONTRASTIVE LEARNING FOR ATTRIBUTED GRAPH CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep attributed graph clustering aims to learn discriminative node representations by jointly exploiting node attributes and graph topology to partition nodes into distinct clusters without label supervision. Although substantial progress has been made in recent years, two key limitations persist. First, propagation operators — whether fixed, as in spectral graph convolutions, or learned, as in attention-based scorers — aggregate neighboring features according to topology alone or according to a scorer with no metric grounding, injecting noise whenever a neighbor's representation is not actually informative for the node's cluster identity. Second, the contrastive objectives typically used to sharpen the resulting representations rely on stochastic graph augmentation and large negative pools, both heuristic, computationally costly, and liable to distort the graph's semantics through arbitrary edge or feature perturbation. We propose OT-GeoCL, an attributed graph clustering method built around a single entropic optimal transport coupling that addresses both limitations at once. Node attributes are mapped onto the unit hypersphere, and a sparse Sinkhorn coupling is computed from the resulting geometric cost, restricted to the graph's edges; this coupling serves simultaneously as a geometry-adaptive propagation operator that down-weights uninformative neighbors by construction, and as the soft positive-pair weighting of a contrastive loss that requires neither augmentation nor heuristic negative sampling. Unlike prior doubly-stochastic propagation schemes, we prove that the resulting training objective admits an exact, non-asymptotic decomposition into an alignment term and a coupling-independent uniformity term, is provably a member of the InfoNCE family with exactly identified limiting cases, and is guaranteed to resist representation collapse via classical spherical potential theory — theoretical guarantees not available to heavier, heuristically-assembled clustering pipelines. Extensive experiments on the twelve-dataset DGCBench benchmark, spanning citation, co-purchase, co-authorship, social, and heterophilous networks, evaluate OT-GeoCL against twelve representative deep graph clustering baselines under a single unified protocol.

1 INTRODUCTION

Clustering is a foundational unsupervised technique for discovering structure in data without label information (Jain et al., 1999). Graph-structured data (pervasive in citation networks, recommendation systems, and molecular structures) adds a second source of signal beyond node attributes: topology. Graph neural networks integrate both (Hamilton et al., 2017; Kipf & Welling, 2017), and deep attributed graph clustering aims to train an encoder that partitions nodes into clusters using structure and attributes jointly, without label supervision. Existing approaches broadly split into two families: graph auto-encoder methods reconstruct topology from a joint structure-attribute encoding, typically paired with a self-training clustering loop (Guo & Dai, 2022), while contrastive methods maximize agreement between augmented views of the graph (You et al., 2020).

Graph neural networks encode both attributes and topology effectively, but two key challenges remain. **(1)** Standard propagation operators, such as fixed spectral GCN kernels or learned attention scorers, aggregate neighboring features without metric grounding. They treat all topological neigh-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bors as informative, regardless of whether a neighbor actually shares the node cluster identity. This mechanism breaks down under heterophily, where topological adjacency and cluster co-membership diverge. **(2)** Contrastive objectives used to sharpen representations against this noise typically rely on stochastic graph augmentation and large negative pools. Both strategies remain heuristic. Augmentation strength is dataset dependent and risks distorting graph semantics, while negative sampling introduces computational overhead without a principled selection rule.

To address these two difficulties jointly, rather than with two separate mechanisms, we propose OT-GeoCL. We map node features onto the unit hypersphere and compute a single sparse entropic Sinkhorn optimal transport coupling $\mathbf{P}$, restricted to the graph's actual edges, from a fixed geometric cost (chordal distance) rather than a learned attention scorer. This coupling addresses difficulty (1) directly: entropic transport concentrates mass on low-cost, geometrically-consistent neighbors and suppresses high-cost ones by construction, before any label or clustering signal is available and irrespective of whether those neighbors are topologically close, so the same mechanism is not tied to a homophily assumption. The same coupling then addresses difficulty (2). Reused as the soft positive pair weighting of a contrastive loss, it removes the need for graph augmentation (positive pairs are given by the real edges of the graph, weighted by transport mass). Additionally, the loss softmax normalization supplies negatives implicitly, with no memory bank or heuristic sampling.

Beyond this architectural design, the single construction resolving (1) and (2) is theoretically well founded. The double stochasticity of $\mathbf{P}$ serves both propagation and loss weighting. This property yields an exact, non asymptotic loss decomposition into alignment and coupling independent uniformity terms (Theorem 1). It also establishes the loss as a precise generalization of the InfoNCE family with exact limiting cases (Proposition 1). Furthermore, it grounds anti collapse behavior in classical spherical potential theory (Corollary 1). These theoretical guarantees are inaccessible to schemes built from heuristic, unrelated components. *Scope.* This work targets the standard attributed graph clustering setting evaluated by the DGCBench benchmark (Wu et al., 2025). This setting spans homophilous citation and co-authorship graphs, heterophilous social graphs, and kNN-constructed graphs from non-graph data.

Contributions.

- We propose a single entropic OT coupling that jointly serves as a geometry-adaptive propagation operator and as the positive-pair weighting of a contrastive loss, addressing noisy propagation and augmentation-dependent contrastive learning with one mechanism that is not tied to a homophily assumption and, as Section 6 shows, is competitive under homophily and strongest under heterophily.
- We show the transport plan's support need not be the observed graph specifically: any fixed, symmetric prior — a k-NN graph over raw features, or curated relational knowledge — can constrain it instead, the graph used here being one instantiation rather than a requirement of the theory.
- We prove that this construction admits an exact, non-asymptotic alignment-uniformity decomposition (Theorem 1), made possible by the coupling's double stochasticity, and that the resulting loss is an exact member of the InfoNCE family, with limiting cases recovering NT-Xent and SupCon-style objectives exactly (Proposition 1), and is guaranteed to resist representation collapse via classical spherical potential theory (Corollary 1).
- We conduct experiments on the twelve-dataset DGCBench benchmark (Wu et al., 2025) against twelve representative baselines under a single unified evaluation protocol.

2 Related Work

To contextualize our proposed framework (OT-GeoCL), we review the relevant literature across three main axes: attributed graph clustering, self-supervised representation learning on graphs, and optimal transport mechanisms. While prior works in these domains often rely on complex architectures, heuristic augmentations, or parametrized attention modules, OT-GeoCL unifies node propagation and contrastive weighting through a single, parameter-free optimal transport mechanism.

Attributed graph clustering. An attributed graph is $G = (V, E, \mathbf{X})$ with node set V, edge set E, and feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$; attributed graph clustering partitions V using both E and $\mathbf{X}$,

without label supervision. DAEGC (Wang et al., 2019) jointly optimizes feature extraction and clustering via a graph attentional auto-encoder; SDCN (Bo et al., 2020) combines an autoencoder with a GCN encoder to counter over-smoothing (Li et al., 2018; Oono & Suzuki, 2020); DFCN (Tu et al., 2021), AGCN (Peng et al., 2021), DAGC (Peng et al., 2023), and DCRN (Liu et al., 2022) add fusion, attention, or correlation-reduction modules. GraphMAE (Hou et al., 2022) reconstructs masked features; AUTOSSL (Jin et al., 2022), ParetoGNN (Ju et al., 2023), and DyFSS (Zhu et al., 2024) fuse pretext tasks via searched, Pareto-balanced, or mixture-of-experts weights. OT-GeoCL differs from this line by using a single mechanism — one entropic transport coupling — for both propagation and contrastive weighting. Evaluating twelve such methods across eleven datasets, Wu et al. (2025) find performance rarely exceeds the dataset's own homophily ratio, corroborating that ungrounded neighbor weighting, not insufficient depth, is a likely bottleneck.

Self-supervised learning. DGI (Velickovic et al., 2019), GraphCL (You et al., 2020), and GCA (Zhu et al., 2021) contrast augmented graph views via InfoNCE-style objectives, requiring augmentation choices and large negative pools; BGRL (Thakoor et al., 2022), CCA-SSG (Zhang et al., 2021), Graph Barlow Twins (Bielak et al., 2022), and AFGRL (Lee et al., 2022) discard negatives instead, via a bootstrap, cross-correlation, or BYOL-style objective. SGC (Wu et al., 2019) shows a fixed, precomputed propagation with logistic regression is competitive with deeper GCNs, at the cost of adaptivity. OT-GeoCL's positive pairs are given directly by the graph's edges, weighted by transport mass: no augmentation, no discovered view, no momentum network.

Optimal transport. Optimal transport (OT) (Villani, 2009) measures distances between distributions by the most cost-efficient transformation of one into another, useful for domain adaptation (Courty et al., 2017) and generative modeling (Tolstikhin et al., 2018); Cuturi (2013) smooth the OT linear program with entropic regularization, computable via Sinkhorn's matrix-scaling algorithm. OT-GeoCL departs from OT's typical two-distribution use by computing a single self-referential coupling over the node embeddings, restricted to the graph's edges, turning Sinkhorn's algorithm into a sparse propagation and contrastive-weighting mechanism. Closer to this construction, Sinkformers (Sander et al., 2022) and SinkGAT (Anonymous, 2022) replace attention's row-stochastic softmax with Sinkhorn's algorithm from a *learned* query/key score, and OT-GNN (Chen et al., 2020) computes Wasserstein distances to learned prototypes; OT-GeoCL instead uses a fixed, geometric pre-Sinkhorn cost and reuses the resulting coupling as the loss's positive-pair weighting (Section 4). Independently, Zass & Shashua (2006) show Normalized Cut and Ratio Cut both seek the doubly-stochastic matrix closest to a raw affinity via the same Sinkhorn–Knopp form we use; Ding et al. (2022) formalize when this projection provably improves cluster separation.

3 Proposed Model

A single entropic optimal transport coupling $\mathbf{P}$, computed from the geometry of the current node embeddings and restricted to the graph's edges, plays two roles at once. *OT-guided propagation* replaces the fixed spectral convolution of a GCN with $\mathbf{P}$, aggregating neighbors through $\mathbf{Z}' = \mathbf{P}\mathbf{Z}$ and suppressing geometrically uninformative ones by construction — addressing difficulty (1). *OT-weighted contrastive* reuses the same $\mathbf{P}$ as the soft positive-pair weighting of a single contrastive loss, requiring neither augmentation nor heuristic negative sampling — addressing difficulty (2).

3.1 Preliminaries and Notation

Let $G = (V, E)$ be an attributed graph with $N = |V|$ nodes, feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$, and adjacency matrix $A \in \{0,1\}^{N \times N}$. We write $\widetilde{E} = E \cup \{(i,i) : i \in V\}$ for the self-loop-augmented edge set and $\widetilde{A} := A + I$, $\mathbf{Z} \in \mathbb{R}^{N \times d}$ for node embeddings with $\mathbf{z}_i \in \mathbb{R}^d$ its i-th row, and $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1\}$ for the hypersphere. The goal is an encoder $f_\theta : \mathbf{X}, G \mapsto \mathbf{Z} \in (\mathbb{S}^{d-1})^N$ well suited to K-means clustering against ground-truth classes, using as little machinery as the task allows.

3.2 From a Topological to a Geometric Graph

A GCN exploits the topology of a graph but is blind to its geometry; OT-GeoCL keeps the same topology but makes the propagation weights a function of embedding geometry itself, without a learned scorer or a second network. GCN's operator $\mathbf{P}_{\mathrm{GCN}} = D^{-1/2}\widetilde{A}D^{-1/2}$ is fixed before training, never querying $\mathbf{Z}$, and is symmetric but generally neither row- nor column-stochastic unless G is regular; GAT (Velickovic et al., 2018) makes the operator dynamic and row-stochastic through a learned score with no relationship to embedding distance, requiring a scorer MLP, and SinkGAT (Anonymous, 2022) doubly-stochastically normalizes this same learned score, inheriting the same extra machinery. OT-GeoCL keeps the topology of GCN and GAT — eligible pairs are exactly $\widetilde{E}$, unchanged throughout training — but replaces the weight's source with a fixed, explicit metric: the chordal distance $c_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$ on $\mathbb{S}^{d-1}$, dynamic and doubly stochastic like SinkGAT but, unlike GAT and SinkGAT, with no learned parameters beyond the base encoder.

3.3 Hyperspherical Embedding and Sparse Entropic Transport Cost

A linear encoder $W \in \mathbb{R}^{F\times d}$ — the only learned weight matrix in the model besides the scalar temperature parameters introduced below — produces

$$\mathbf{h}_i = W\mathbf{x}_i, \qquad \mathbf{z}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2}, \qquad i = 1,\dots,N, \tag{1}$$

constraining every embedding to $\mathbb{S}^{d-1}$. For every $(i,j) \in \widetilde{E}$, the transport cost is read directly off the hyperspherical geometry, with no separate cost network:

$$c_{ij} \;=\; \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \;=\; 2 - 2\langle \mathbf{z}_i, \mathbf{z}_j\rangle, \tag{2}$$

i.e. twice one minus the cosine similarity. A single learnable scalar $\sigma > 0$ (inverse temperature) instantiates a Gibbs kernel,

$$K_{ij} \;=\; \exp\Big(-\frac{c_{ij}}{2\sigma^2}\Big), \qquad (i,j) \in \widetilde{E}, \tag{3}$$

materialized only on the support of $\widetilde{E}$, never as a dense $N \times N$ matrix.

3.4 Sparse Log-Domain Sinkhorn Coupling

A fixed, reliable prior decides who can be a neighbor; geometry decides how much each neighbor matters. Entropic OT (Cuturi, 2013) usually couples two point clouds with no restriction on which pairs may be matched: the transport polytope $\Pi(\mathbf{1},\mathbf{1})$ is dense. We instead solve a *support-constrained* transport problem, restricting the feasible set to the intersection of $\Pi(\mathbf{1},\mathbf{1})$ with a fixed, symmetric prior $\mathcal{S} \subseteq V \times V$,

$$\min_{\mathbf{P}\in\Pi(\mathbf{1},\mathbf{1})\,\cap\,\mathrm{supp}(\mathcal{S})} \;\sum_{(i,j)\in\mathcal{S}} P_{ij}\,c_{ij} \;-\; \varepsilon\,H(\mathbf{P}), \tag{4}$$

with H the entropy and $\varepsilon = \sigma^2$. Neither Theorem 1 nor Proposition 1 requires $\mathcal{S}$ to be the observed graph specifically: they only need $\mathcal{S}$ to be fixed before training, symmetric, and to contain the diagonal. Throughout this paper $\mathcal{S} = \widetilde{E}$, the graph's own edges, but $\mathcal{S}$ could equally be a k-NN graph over raw features or an externally curated relation whenever more trustworthy than the observed edges themselves. Unlike structured-OT formulations that softly penalize undesired couplings through the cost (Alvarez-Melis et al., 2018), the constraint $\mathrm{supp}(\mathbf{P}) \subseteq \mathcal{S}$ is hard and combinatorial: a pair absent from $\mathcal{S}$ is never assigned mass, never assigned a cost, at any value of σ.

This restricted problem is solved by log-domain Sinkhorn–Knopp iterations on dual potentials $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, initialized at $\mathbf{0}$ and updated for T iterations:

$$u_i^{(t+1)} = -\,\mathrm{LSE}_{j:(i,j)\in\widetilde{E}}\Big(\log K_{ij} + v_j^{(t)}\Big), \tag{5}$$

$$v_j^{(t+1)} = -\,\mathrm{LSE}_{i:(i,j)\in\widetilde{E}}\Big(\log K_{ij} + u_i^{(t+1)}\Big), \tag{6}$$

with LSE a sparse, scatter-based log-sum-exp over each node's neighbors, recovering

$$P_{ij} \;=\; \exp\Big(\log K_{ij} + u_i + v_j\Big), \qquad (i,j) \in \widetilde{E}, \tag{7}$$

after the final iteration. Every sum in Eqs. (5)–(6) runs over $\{j : (i,j) \in \widetilde{E}\}$ rather than all N nodes, keeping the computational complexity of each iteration at $\mathcal{O}(|E|)$; T and σ are the only knobs introduced by this step.

Remark 1 (Why constraining the support, not just the cost, matters)**.** *Solving Eq.* (4) *over* $\mathrm{supp}(\widetilde{E})$ *rather than the full transport polytope has two consequences beyond the* $\mathcal{O}(|E|)$ *complexity of Section 5.* Interpretability: *since the support never changes,* $\pi_i(j) := P_{ij}$ *is a genuine probability distribution over* i*'s actual graph neighbors.* Well-posedness: *Sinkhorn–Knopp iterations converge to a doubly-stochastic limit iff the matrix has at least one positive diagonal (Sinkhorn & Knopp, 1967); the self-loops in* $\widetilde{E}$ *guarantee this, with uniqueness additionally requiring full indecomposability, satisfied whenever the graph is connected. Both guarantees require* $\mathcal{S}$ *to be symmetric; a general prior must be symmetrized (*$\mathcal{S} \cup \mathcal{S}^\top$*) first, or double stochasticity, and with it Theorem 1, no longer holds.*

3.5 OT-Guided Aggregation and Contrastive Objective

Given $\mathbf{P}$, representations update by a single mass-preserving diffusion step,

$$\mathbf{z}'_i \;=\; \sum_{j:(i,j)\in\widetilde{E}} P_{ij}\,\mathbf{z}_j, \qquad \mathbf{Z}' \;=\; \mathbf{P}\,\mathbf{Z}, \tag{8}$$

followed by re-projection onto the hypersphere, $\mathbf{Z}_{\mathrm{out}} = \mathbf{Z}'/\|\mathbf{Z}'\|_{2,\mathrm{row}}$ — structurally a single, doubly-stochastic, support-restricted step of blurring mean-shift (Carreira-Perpinan, 2006).

Rather than contrasting augmented views, OT-GeoCL contrasts each node against all others, using $\mathbf{P}$ to decide how strongly each edge should be pulled together. With $\mathbf{Z}$ the propagated, re-normalized embeddings,

$$S_{ik} = \frac{\langle \mathbf{z}_i, \mathbf{z}_k \rangle}{\tau}, \qquad \log p_{i|k} = \log\left(\frac{\exp(S_{ik})}{\sum_{k'=1}^{N}\exp(S_{ik'})}\right), \tag{9}$$

Edges $(i,j) \in \widetilde{E}$ are soft positive pairs weighted by normalized transport mass $\bar{P}_{ij} = P_{ij}/\sum_{(i',j')\in\widetilde{E}} P_{i'j'}$; OT-GeoCL's entire training objective is the single term

$$\mathcal{L}_{\mathrm{OT}} \;=\; -\frac{1}{2}\sum_{(i,j)\in\widetilde{E}} \bar{P}_{ij}\Big(\log p_{i|j} + \log p_{j|i}\Big). \tag{10}$$

There is no second loss term, no clustering-alignment objective, and no auxiliary supervision signal. Local attraction along OT-coupled edges, proportional to $\bar{P}_{ij}$, is counterbalanced by the softmax normalization's implicit repulsion against the full node population — every non-neighbor an implicit negative, with no memory bank or augmentation pipeline required to produce them.

Proposition 1 ($\mathcal{L}_{\mathrm{OT}}$ generalizes InfoNCE)**.** *Let* $\pi_i(j) := P_{ij}$ *for* $j \in \mathcal{N}(i) := \{j : (i,j) \in \widetilde{E}\}$*; row-stochasticity of* $\mathbf{P}$ *(Eq.* (5)*) guarantees* $\sum_{j\in\mathcal{N}(i)}\pi_i(j) = 1$*, so* $\pi_i(\cdot)$ *is a genuine per-anchor probability distribution over neighbors. Then:* (a) Exact recovery. *If* $\widetilde{E}$ *restricts to a perfect matching (one positive per anchor, as under paired-view augmentation),* $\mathcal{L}_{\mathrm{OT}}$ *equals the symmetric NT-Xent/SimCLR loss exactly, independent of* σ. (b) High-temperature limit. *On a* d*-regular graph, as* $\sigma \to \infty$*,* $\mathcal{L}_{\mathrm{OT}}$ *converges to the uniform multi-positive (SupCon-style) loss.* (c) General case. *For finite* σ *and an arbitrary graph,* $\pi_i(\cdot)$ *is a differentiable, geometry-adaptive relaxation of the fixed, combinatorial weighting used in (a) and (b).*

Proof in Appendix A.

Algorithm 1 is the complete forward pass: one encoder, one coupling, one loss, no pretraining or fine-tuning phases. $\mathcal{L}_{\mathrm{OT}}$ is differentiable in $\theta = \{W, \sigma\}$ end to end, including through the T Sinkhorn iterations; training repeats Algorithm 1 once per epoch, backpropagating with Adam.

4 Theoretical Discussion: An Exact Alignment-Uniformity Decomposition

Wang & Isola (2020) show that, as $N \to \infty$, a broad family of contrastive losses converges to a sum of an *alignment* term and a *uniformity* term, and that minimizing the latter is what prevents repre-

Algorithm 1 OT-GeoCL forward pass (single layer, single loss)

Require: Features $\mathbf{X}$, edges E, encoder W, scale σ, iterations T

$\widetilde{E} \leftarrow E \cup \{(i,i)\}_{i \in V}$
$\mathbf{Z} \leftarrow \mathrm{RowNormalize}(W\mathbf{X})$ ▷ Eq. (1)
for $(i,j) \in \widetilde{E}$ **do**
    $c_{ij} \leftarrow 2 - 2\langle \mathbf{z}_i, \mathbf{z}_j \rangle$; $\log K_{ij} \leftarrow -c_{ij}/2\sigma^2$
end for
$\mathbf{u}, \mathbf{v} \leftarrow \mathbf{0}, \mathbf{0}$
for $t = 1, \ldots, T$ **do**
    update $\mathbf{u}$ via Eq. (5); update $\mathbf{v}$ via Eq. (6)
end for
for $(i,j) \in \widetilde{E}$ **do**
    $P_{ij} \leftarrow \exp(\log K_{ij} + u_i + v_j)$
end for
$\mathbf{Z}' \leftarrow \mathbf{P}\mathbf{Z}$; $\mathbf{Z}_{\mathrm{out}} \leftarrow \mathrm{RowNormalize}(\mathbf{Z}')$
compute $\mathcal{L}_{\mathrm{OT}}$ via Eq. (10) using $\mathbf{Z}_{\mathrm{out}}$ and $\mathbf{P}$
return $\mathbf{Z}_{\mathrm{out}}, \mathcal{L}_{\mathrm{OT}}$ ▷ $\theta = \{W, \sigma\}$ updated externally by gradient descent on $\mathcal{L}_{\mathrm{OT}}$

sentation collapse. The decomposition below holds *exactly*, at any finite N, as a direct consequence of $\mathbf{P}$'s double stochasticity.

Lemma 1 (Symmetry of the coupling)**.** *Since $c_{ij} = c_{ji}$ (Eq. (2) is a genuine metric) and both target marginals in Eq. (4) equal $\mathbf{1}$, the Sinkhorn fixed point satisfies $u = v$, hence $P_{ij} = P_{ji}$. Combined with row-stochasticity, $\mathbf{P}$ is therefore also column-stochastic: $\sum_i P_{ij} = 1$ for every j.*

Theorem 1 (Exact alignment-uniformity decomposition)**.** *Let $\log Z_i := \log \sum_{k=1}^{N} \exp(\langle \mathbf{z}_i, \mathbf{z}_k \rangle / \tau)$. Then, exactly (not asymptotically),*

$$\mathcal{L}_{\mathrm{OT}} = \mathcal{L}_{\mathrm{align}}^{\mathrm{OT}} + \mathcal{L}_{\mathrm{unif}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{align}}^{\mathrm{OT}} := -\frac{1}{N\tau} \sum_{i=1}^{N} \sum_{j:(i,j)\in\widetilde{E}} P_{ij} \langle \mathbf{z}_i, \mathbf{z}_j \rangle$ and $\mathcal{L}_{\mathrm{unif}} := \frac{1}{N} \sum_{i=1}^{N} \log Z_i$, and $\mathcal{L}_{\mathrm{unif}}$ does not depend on $\mathbf{P}$, hence not on the graph, σ, or T, at all.

Proof in Appendix B.

Corollary 1 (Anti-collapse via spherical potential theory)**.** *In the large-N limit, $\mathcal{L}_{\mathrm{unif}}$ converges to the log-partition energy of the empirical measure μ of $\{\mathbf{z}_i\}$ on $\mathbb{S}^{d-1}$. Since $\exp(\langle \cdot, \cdot \rangle / \tau)$ is a completely monotonic function of the chordal distance, Cohn & Kumar (2007) show this energy is uniquely minimized, among all probability measures on $\mathbb{S}^{d-1}$, by the uniform (surface) measure, and uniquely maximized by any Dirac measure. Minimizing $\mathcal{L}_{\mathrm{unif}}$ therefore drives $\{\mathbf{z}_i\}$ away from collapse and toward a spread-out configuration, without any auxiliary anti-collapse term.*

Theorem 1 is, to our knowledge, the first *exact* (rather than asymptotic) alignment-uniformity decomposition in the contrastive learning literature, possible only because $\mathbf{P}$'s double stochasticity eliminates the graph dependence from $\mathcal{L}_{\mathrm{unif}}$ at any finite N; row-stochasticity alone (used for Proposition 1) does not suffice, since a row-stochastic-only coupling (GAT, or SinkGAT's un-symmetrized variant) would leave a residual, graph-dependent term in its place. Because $\mathcal{L}_{\mathrm{OT}}$ decomposes into an alignment term over exactly the positive-pair graph used by HaoChen et al. (2021)'s provable spectral-recovery result and a $\mathbf{P}$-independent uniformity term, we conjecture an analogous recovery guarantee holds for $\mathcal{L}_{\mathrm{OT}}$ under a connectivity condition stated on $\mathbf{P}$ rather than raw adjacency; establishing this is left to future work.

5 Complexity Analysis

Two components of Algorithm 1 scale differently. The coupling itself — Sinkhorn iterations and the propagation step $\mathbf{P}\mathbf{Z}$ — is $\mathcal{O}(T|E|)$, genuinely linear in the number of edges: no dense $N \times N$ matrix is ever instantiated (Remark 1). The contrastive loss's softmax normalization, $\log Z_i = \log \sum_{k=1}^{N} \exp(\langle \mathbf{z}_i, \mathbf{z}_k \rangle / \tau)$, is a separate $\mathcal{O}(N^2)$ cost per epoch, intrinsic to full-softmax InfoNCE-style objectives and shared with SimCLR. This term ran unmodified, with no chunking, on every benchmark in Table 1, up to $N = 22{,}662$ (Roman-Empire); at larger scale it would become binding,

and a chunked, blockwise log-sum-exp implementation of $\log Z_i$ is available to bound peak memory at added wall-clock cost, with no change to $\mathcal{L}_{\mathrm{OT}}$, though none of the datasets evaluated here required it. Overall per-epoch cost is accordingly $\mathcal{O}(T|E| + N^2)$.

Sinkhorn's iteration converges linearly in Hilbert's projective metric (Cuturi, 2013); for $\mathbf{P}$ restricted to $\tilde{E}$, the rate plausibly depends on the graph's algebraic connectivity rather than N alone, consistent with $T = 5$ sufficing across the well-connected benchmarks of Table 1.

6 Experiments

We evaluate OT-GeoCL on the DGCBench benchmark (Wu et al., 2025), comparing against twelve representative deep graph clustering baselines under a single, unified protocol.

6.1 Baselines

Following the unified benchmark of Wu et al. (2025), we compare against twelve state-of-the-art deep graph clustering methods spanning three paradigms. **Reconstructive:** GAE / GAE-SSC (GAE_S) (Kipf & Welling, 2016), DAEGC (Wang et al., 2019), SDCN (Bo et al., 2020), DFCN (Tu et al., 2021) — autoencoder- or GCN-based reconstruction, with GAE_S, SDCN, and DFCN adding self-supervised clustering or fusion modules. **Predictive:** DGCluster (DGClu) (Bhowmick et al., 2024) trains a GNN to predict node similarity guided by modularity maximization, followed by Birch clustering. **Contrastive:** AGC-DRR (Gong et al., 2022) (adaptive edge dropout), DCRN (Liu et al., 2022) (dual correlation reduction), HSAN (Liu et al., 2023) (hardness-weighted pairs), CCGC (Yang et al., 2023) (confidence-based positives/negatives), MAGI (Liu et al., 2024b) (modularity pretext task), NS4GC (Liu et al., 2024a) (cross-view similarity). All baseline figures are as published in Wu et al. (2025); OT-GeoCL is implemented in Python/PyTorch.

6.2 Datasets

Following Wu et al. (2025), we evaluate on eleven datasets spanning three structural regimes: six **homogeneous**, high-homophily graphs (Wiki, Cora, Citeseer, ACM, DBLP, PubMed), three **heterogeneous**, low-homophily graphs (BlogCatalog, Flickr, Roman-Empire), and two k**NN-constructed** graphs from non-graph data (USPS, HHAR, $k = 3$), spanning a homophily ratio from 0.05 to 0.98. Table 1 summarizes their statistics, taken from Wu et al. (2025). We do not report ogbn-arXiv, DGCBench's twelfth dataset, as it was not part of the runs used to produce our tables.

Table 1: Dataset statistics, following Wu et al. (2025). Avg. Degree is $\bar{d} = 2|E|/N$.

Type	Dataset	N	$\lvert E\rvert$	F	Classes	Homoph.	$\bar{d}$
Homog.	Wiki	2,405	17,981	4,973	17	0.71	14.95
	Cora	2,708	5,429	1,433	7	0.81	4.01
	ACM	3,025	13,128	1,870	3	0.82	8.68
	Citeseer	3,327	9,104	3,703	6	0.74	5.47
	DBLP	4,057	3,528	334	4	0.80	1.74
	PubMed	19,717	88,648	500	3	0.80	8.99
Heterog.	BlogCatalog	5,196	343,486	8,189	6	0.40	132.21
	Flickr	7,575	479,476	12,047	9	0.24	126.59
	Roman-Empire	22,662	65,854	300	18	0.05	5.81
kNN	USPS	9,298	27,894	256	10	0.98	6.00
	HHAR	10,299	30,897	561	6	0.95	6.00

6.3 Experimental Protocol

Following Wu et al. (2025)'s standardized pipeline, clustering assignments are obtained by K-means ($K =$ number of classes) on the learned embeddings, with Accuracy (ACC) computed after Hungarian-algorithm label matching and Normalized Mutual Information (NMI) computed directly on the resulting partitions; ACC is the fraction of correctly assigned labels under the optimal cluster-to-label matching, and NMI measures the information shared between predicted and ground-truth partitions, normalized to $[0, 1]$ and permutation-invariant. Each experiment is repeated ten times with a distinct random seed; we report the mean, using the same protocol for OT-GeoCL as for the twelve baselines. Node embeddings are produced by the single OT-GeoCL layer of Algorithm 1, with output dimension $d = 128$, $T = 5$ Sinkhorn iterations, and Adam optimization; τ, learning

Table 2: Clustering performance (%) on the six homogeneous, high-homophily DGCBench benchmarks. Baseline rows are values as reported (final, algorithm-termination results) in Wu et al. (2025), not re-run by us. The OT-GeoCL row reports results under this same protocol (Appendix C details per-dataset hyperparameters); "–" marks GPU memory overflow, as reported by Wu et al. (2025). **Bold** marks the best result per column, underline the second-best.

	Wiki		Cora		ACM		Citeseer		DBLP		PubMed	
Method	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE (Kipf & Welling, 2016)	44.44	44.85	65.56	51.48	87.96	62.05	58.21	32.10	47.00	19.07	64.46	23.12
GAE_S (Kipf & Welling, 2016)	39.13	34.76	70.14	53.39	85.07	60.66	56.44	30.70	51.65	21.70	66.36	24.69
DAEGC (Wang et al., 2019)	39.25	33.37	66.81	48.54	84.36	51.81	60.19	33.45	61.40	23.64	65.33	24.32
SDCN (Bo et al., 2020)	18.66	03.03	50.82	31.75	88.15	63.73	53.46	30.19	61.93	31.76	55.59	19.60
DFCN (Tu et al., 2021)	43.72	42.99	48.85	40.58	89.96	67.20	55.50	36.30	71.48	38.70	47.65	07.34
DGClu (Bhowmick et al., 2024)	**55.17**	45.74	58.41	48.27	77.37	52.17	47.04	24.18	35.90	05.86	59.38	20.79
AGC-DRR (Gong et al., 2022)	45.19	40.40	66.81	52.11	90.71	70.86	65.79	41.52	**79.76**	**48.87**	62.32	25.23
DCRN (Liu et al., 2022)	34.72	27.00	64.71	52.49	91.83	**71.66**	68.31	44.11	70.53	40.78	–	–
HSAN (Liu et al., 2023)	45.55	45.34	63.90	52.43	53.46	23.38	53.47	34.83	66.74	38.85	–	–
CCGC (Yang et al., 2023)	46.82	45.54	61.70	47.65	70.62	48.50	60.11	36.05	33.82	05.05	45.92	09.23
MAGI (Liu et al., 2024b)	29.70	27.26	73.24	55.75	83.27	53.45	68.19	43.34	67.51	37.70	61.81	19.12
NS4GC (Liu et al., 2024a)	44.54	42.63	70.48	56.83	79.17	48.71	67.21	43.32	77.40	45.44	68.92	31.28
OT-GeoCL (ours)	55.09	**52.77**	**76.20**	**60.09**	**91.89**	71.46	**71.34**	**46.04**	79.13	48.00	**70.51**	**35.04**

Table 3: Clustering performance (%) on the two kNN-constructed benchmarks and the three heterogeneous, low-homophily benchmarks. Baseline rows are values as reported in Wu et al. (2025), not re-run by us; "–" marks a result not reported (GPU memory overflow, or, for HSAN, a method restricted to graph-structured input) by Wu et al. (2025). **Bold** marks the best result per column, underline the second-best.

	kNN-constructed				*Heterogeneous, low-homophily*					
	USPS (k=3)		HHAR (k=3)		BlogCatalog		Flickr		Roman-Empire	
Method	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
GAE (Kipf & Welling, 2016)	58.08	52.51	44.64	43.96	46.58	30.16	27.28	15.96	16.12	09.66
GAE_S (Kipf & Welling, 2016)	66.20	59.08	42.55	50.17	45.60	28.25	22.63	12.56	15.88	08.67
DAEGC (Wang et al., 2019)	67.96	66.18	58.00	57.48	40.74	23.63	26.43	12.43	–	–
SDCN (Bo et al., 2020)	**73.20**	74.86	69.21	70.48	24.04	07.57	17.30	06.49	14.61	01.44
DFCN (Tu et al., 2021)	72.95	**75.06**	67.18	73.01	37.85	26.79	31.13	18.58	12.02	06.24
DGClu (Bhowmick et al., 2024)	71.28	69.08	69.15	72.70	45.06	27.24	23.17	11.30	11.54	03.74
AGC-DRR (Gong et al., 2022)	67.33	67.91	65.20	60.68	75.68	65.48	31.39	28.93	–	–
DCRN (Liu et al., 2022)	26.86	22.29	36.63	46.96	56.68	42.48	18.36	08.94	–	–
HSAN (Liu et al., 2023)	–	–	–	–	45.40	30.98	–	–	–	–
CCGC (Yang et al., 2023)	38.82	36.20	39.38	41.93	29.42	10.95	18.25	07.53	16.26	09.32
MAGI (Liu et al., 2024b)	57.03	53.98	45.04	45.93	41.72	24.44	31.21	16.06	17.90	06.94
NS4GC (Liu et al., 2024a)	69.57	66.88	60.56	55.88	36.36	20.77	31.51	01.84	14.42	08.08
OT-GeoCL (ours)	70.71	68.79	**79.51**	**74.07**	**93.89**	**83.48**	**74.09**	**62.08**	**31.37**	**37.23**

rate, weight decay, and training length are tuned per dataset, and σ's per-dataset *initial* value is set manually but its trained value is learned by gradient descent alongside the encoder (Eq. (3)). Baseline hyperparameters follow the original papers or public source-code defaults, exactly as in Wu et al. (2025).

6.4 Analysis of the Cluster Results

Tables 2 and 3 report the baseline comparison together with OT-GeoCL. The clearest pattern is a regime split. On the three benchmarks Table 1 labels **heterogeneous, low-homophily** — BlogCatalog, Flickr, and Roman-Empire — OT-GeoCL is the best method on *both* ACC and NMI, without exception, and by wide margins over the best baseline on each dataset: BlogCatalog ($+18.21$ ACC, $+18.00$ NMI, both over AGC-DRR), Flickr ($+42.58$ ACC over NS4GC, $+33.15$ NMI over AGC-DRR), and Roman-Empire ($+13.47$ ACC over MAGI, $+27.57$ NMI over GAE). This revises, rather than confirms, Wu et al. (2025)'s finding that every method's ACC on Roman-Empire (homophily 0.05) falls below 22%: OT-GeoCL reaches 31.37%, nine points past that ceiling, with an NMI (37.23) close to four times the best baseline's (9.66). The two kNN-constructed graphs split: OT-GeoCL is best on HHAR ($+10.30$ ACC, $+1.06$ NMI) but trails SDCN on USPS on both metrics (-2.49 ACC, -6.27 NMI). On the six **homogeneous, high-homophily** benchmarks of Table 2, the picture is more mixed and the margins are narrower: OT-GeoCL is best on both metrics for Cora, Citeseer, and PubMed; splits with the best baseline on Wiki (best NMI by a wide margin, $+7.03$, and essentially tied on ACC, -0.08) and ACM (best ACC by 0.06 points, second-best NMI, -0.20); and trails AGC-DRR on DBLP by under a point on each metric (-0.63 ACC, -0.87 NMI).

Read together, these results qualify Wu et al. (2025)'s homophily-bottleneck narrative rather than simply confirming it: OT-GeoCL's advantage over every baseline is largest precisely where topology is least trustworthy — Roman-Empire (homophily 0.05), Flickr (0.24), and BlogCatalog (0.40) — and narrower or negative on the high-homophily graphs (Wiki, ACM, DBLP), where most edges are already informative. This is consistent with the coupling acting as a geometry-driven *edge selector*: because the transport cost $c_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$ (Section 3) depends only on embedding geometry, entropic Sinkhorn assigns label-crossing edges vanishing coupling mass whenever their endpoints are geometrically dissimilar, regardless of the edge's mere topological presence — a soft pruning of the graph's own topology, rather than the uniform averaging $\mathbf{P}_{\mathrm{GCN}} = D^{-1/2}\widetilde{A}D^{-1/2}$ performs over the same edges. The exploratory heterophilous ablation of Appendix C gives a quantitative handle on this: replacing $\mathbf{P}_{\mathrm{GCN}}$ with the full geometric coupling under $\mathbf{P}$-weighted loss improves ACC by $+13.75$ points on average, against only $+2.06$ points under uniform loss, so propagation and loss-weighting reinforce each other far more sharply under heterophily than the homophilous pattern expected from the DGCBench ablation of Appendix C.3. We report this as the empirical pattern the tables and this ablation support, not as established by a controlled intervention isolating geometry from topology.

We emphasize the scope of this claim: none of the eleven DGCBench baselines compared in Tables 2–3 is purpose-built for heterophily in the way that, e.g., geometry-based neighborhood redefinition (Pei et al., 2020) is; these twelve are homophily-oriented clustering methods evaluated as-is on low-homophily graphs by Wu et al. (2025). The results here therefore show that OT-GeoCL degrades markedly less than these methods under low homophily, not that it has been benchmarked against the dedicated heterophilous-GNN literature, which we leave to future comparison. An ablation on the same DGCBench datasets, an exploratory heterophilous extension, an embedding-dimension sensitivity study, and qualitative analyses are reported in Appendix C.

6.5 Statistical Significance across the Main Benchmark

The comparisons above are rank-based summaries; we complement them with a Friedman test and post-hoc Nemenyi test (Demsar, 2006) on the NMI values of Tables 2–3, over the nine methods with complete data on the eleven datasets that admit a well-defined block design (Appendix C.4 gives the exclusion criteria and full pairwise results). The Friedman test rejects the null hypothesis of equal performance ($\chi_F^2 = 31.01$, $df = 8$, $p = 1.4 \times 10^{-4}$); OT-GeoCL obtains the best average rank (1.27 of 9) and is significantly better under Nemenyi ($\alpha = 0.05$) than seven of the eight baselines, though not significantly separated from its closest competitor by rank (DFCN), a power limitation of testing over eleven datasets rather than evidence of equivalence.

7 Conclusion

This paper introduces OT-GeoCL, an attributed graph clustering method centered on a single entropic optimal transport coupling. This coupling serves both as a geometry adaptive propagation operator and as the positive pair weighting for a contrastive loss. Thanks to its double stochasticity, this unified construction enjoys strong theoretical guarantees that modular architectures lack: an exact, non asymptotic alignment uniformity decomposition (Theorem 1), an exact connection to the InfoNCE family (Proposition 1), and proof of anti collapse via spherical potential theory (Corollary 1). Evaluated on eleven datasets of the DGCBench benchmark against twelve representative baselines under a single unified protocol, OT-GeoCL is competitive with or ahead of the strongest baseline on half of the homogeneous, high-homophily benchmarks, and is the uniformly best method, on both accuracy and NMI, across all three heterogeneous, low-homophily benchmarks, including a nine-point accuracy gain on Roman-Empire, the lowest-homophily graph in the suite, consistent with a coupling that adapts to embedding geometry rather than to topology alone. Reusing $\mathbf{P}$ for both propagation and loss weighting is central to this design, but the heterophilous ablation of Appendix C suggests it contributes unevenly across regimes, reinforcing more sharply under heterophily; confirming this on the homophilous DGCBench datasets, studying which structural properties govern the balance, and extending the coupling's cost to heterophilous edge patterns are left to future work.

Reproducibility Statement

$\mathbf{P}$ and the objective are given in closed form in Eqs. (1)–(10) and Algorithm 1; proofs are in Appendix A–B. Dataset sources follow Wu et al. (2025) (Table 1); hyperparameters and the full procedure (Algorithm 2) are in Appendix C.

Ethics Statement

This work introduces a general-purpose unsupervised method evaluated on standard, public benchmarks; it involves no human subjects or personally identifiable data, and raises no ethical concerns beyond those of graph representation learning generally.

AI Use Statement

Generative AI tools assisted with drafting this text and with implementing/debugging code, but not with generating results, fabricating citations, or writing the proofs of Theorem 1 or Proposition 1; the authors derived and verified these and take full responsibility for the final content.

References

David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. Structured optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Anonymous. SinkGAT: Doubly-stochastic graph attention, 2022. Preprint, under review. Author list not independently confirmed at the time of writing; cited by title only.

Aritra Bhowmick, Mert Kosan, Zexi Huang, Ambuj Singh, and Sourav Medya. DGCluster: A neural framework for attributed graph clustering via modularity maximization. In *AAAI Conference on Artificial Intelligence*, 2024.

Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V. Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *Knowledge-Based Systems*, 256:109631, 2022.

Deyu Bo, Xiao Wang, Chuan Shi, Meiqi Zhu, Emiao Lu, and Peng Cui. Structural deep clustering network. In *The Web Conference (WWW)*, 2020.

Miguel A. Carreira-Perpinan. Fast nonparametric clustering with gaussian blurring mean-shift. In *International Conference on Machine Learning (ICML)*, 2006.

Benson Chen, Gary Becigneul, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Optimal transport graph neural networks. *arXiv preprint arXiv:2006.04804*, 2020.

Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9), 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Tianjiao Ding, Derek Lim, Rene Vidal, and Benjamin D. Haeffele. Understanding doubly stochastic clustering. In *International Conference on Machine Learning (ICML)*, 2022.

Lei Gong, Sihang Zhou, Wenxuan Tu, and Xinwang Liu. Attributed graph clustering with dual redundancy reduction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Lin Guo and Qun Dai. Graph clustering via variational graph embedding. *Pattern Recognition*, 122: 108334, 2022.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

Wei Jin, Xiaorui Liu, Xiangyu Zhao, Yao Ma, Neil Shah, and Jiliang Tang. Automated self-supervised learning for graphs. In *International Conference on Learning Representations (ICLR)*, 2022.

Mingxuan Ju, Tong Zhao, Qianlong Wen, Wenhao Yu, Neil Shah, Yanfang Ye, and Chuxu Zhang. Multi-task self-supervised graph neural networks enable stronger task generalization. In *International Conference on Learning Representations (ICLR)*, 2023.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Namkyeong Lee, Junseok Lee, and Chanyoung Park. Augmentation-free self-supervised learning on graphs. In *AAAI Conference on Artificial Intelligence*, 2022.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.

Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *AAAI Conference on Artificial Intelligence*, 2022.

Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *AAAI Conference on Artificial Intelligence*, 2023.

Yue Liu, Xihong Gao, Tianyu He, Tong Zheng, Jieyu Zhao, and Hongzhi Yin. Reliable node similarity matrix guided contrastive graph clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(12), 2024a.

Yue Liu, Jun Li, Yihao Chen, Ruiyi Wu, Ershen Wang, Jun Zhou, Shuo Tian, Shu Shen, Xingtian Fu, Chao Meng, et al. Revisiting modularity maximization for graph clustering: A contrastive learning perspective. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2024b.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Zhihao Peng, Hui Liu, Yuheng Jia, and Junhui Hou. Attention-driven graph clustering network. In *ACM International Conference on Multimedia*, 2021.

Zhihao Peng, Hui Liu, Yuheng Jia, and Junhui Hou. Deep attention-guided graph clustering with dual self-supervision. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7), 2023.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations (ICLR)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyre. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Remi Munos, Petar Velickovic, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2022.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Scholkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *AAAI Conference on Artificial Intelligence*, 2021.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Petar Velickovic, William Fedus, William L. Hamilton, Pietro Lio, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations (ICLR)*, 2019.

Cedric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Chun Wang, Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, and Chengqi Zhang. Attributed graph clustering: A deep attentional embedding approach. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

Binbin Wu, Yifang Liu, Qincheng Tan, Xiao Liu, Wenqi Du, Jing Wang, and Ge Yu. DGCBench: A deep graph clustering benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning (ICML)*, 2019.

Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu. Cluster-guided contrastive graph clustering network. In *AAAI Conference on Artificial Intelligence*, 2023.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2006.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S. Yu. From canonical correlation analysis to self-supervised graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Pengfei Zhu, Qian Wang, Yu Wang, Jialu Li, and Qinghua Hu. Every node is different: Dynamically fusing self-supervised tasks for attributed graph clustering. In *AAAI Conference on Artificial Intelligence*, 2024.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *The Web Conference (WWW)*, 2021.

A Proof of Proposition 1

Step 1: row-stochasticity turns edge weights into a per-anchor distribution. By construction (Eq. (5)), immediately after every u-update,

$$\sum_{j:(i,j)\in\widetilde{E}} P_{ij} = 1 \quad \text{for every } i, \tag{12}$$

so each row of $\mathbf{P}$ is already a probability distribution over i's neighbors: $\pi_i(j) := P_{ij}$. Since $\sum_{(i,j)\in\widetilde{E}} P_{ij} = N$ (one unit of mass per row), the normalized edge weight in Eq. (10) decomposes as

$$\bar{P}_{ij} = \frac{1}{N}\pi_i(j), \tag{13}$$

the usual $1/N$ anchor average composed with a per-anchor soft assignment. Substituting into Eq. (10) gives the equivalent per-anchor form

$$\mathcal{L}_{\mathrm{OT}} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} \pi_i(j)\Big(\log p_{i|j} + \log p_{j|i}\Big). \tag{14}$$

Step 2: this is the shape of a weighted multi-positive InfoNCE loss. The single-positive, symmetrized NT-Xent loss used in SimCLR-style frameworks (You et al., 2020) is

$$\mathcal{L}_{\text{NT-Xent}} = \frac{1}{2N}\sum_{i=1}^{N}\big(\ell(i,\mathrm{pos}(i)) + \ell(\mathrm{pos}(i),i)\big), \tag{15}$$

with $\ell(i,j) := -\log p_{i|j}$ and $\mathrm{pos}(i)$ the unique positive of i. Its multi-positive generalization is

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{N}\sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log p_{i|p}, \tag{16}$$

with positive set $P(i)$ uniformly weighted. Both are instances of the generic template

$$\mathcal{L}_{w} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j\in P(i)} w_i(j)\big(\log p_{i|j} + \log p_{j|i}\big), \tag{17}$$

with $w_i(j) = \mathbb{1}[j = \mathrm{pos}(i)]$ giving Eq. (15) and $w_i(j) = 1/|P(i)|$ giving Eq. (16). Equation (14) is exactly this template with $P(i) = \mathcal{N}(i)$ and $w_i(j) = \pi_i(j)$.

Step 3: the three cases. (a) With one edge per anchor, Eq. (12) forces $\pi_i(\mathrm{pos}(i)) = 1$; substituting into Eq. (14) yields Eq. (15) exactly. (b) As $\sigma \to \infty$, $\log K_{ij} \to 0$ on $\widetilde{E}$; the entropic problem of Eq. (4) degenerates to the doubly-stochastic scaling of $\widetilde{E}$'s binary adjacency matrix, which on a d-regular graph is uniquely $1/d$ on every edge, giving $\pi_i(j) = 1/d = 1/|\mathcal{N}(i)|$ and hence Eq. (16). (c) follows directly from Eqs. (2)–(7) defining $\pi_i(\cdot)$ as a function of $\mathbf{Z}$, hence of the encoder parameters W. ■

Remark 2 (Uniform weighting is not, by itself, InfoNCE)**.** *A natural but incorrect shortcut is to say InfoNCE is recovered whenever $\pi_i(\cdot)$ is uniform. This holds only when the positive set has cardinality one (case (a)). If instead $\pi_i(j) = 1/N$ uniformly over all N nodes (the complete graph, every node a positive of every other), Eq.* (14) *collapses by symmetry to $\mathcal{L}_{\mathrm{OT}} = -\frac{1}{N^2}\sum_{i,j}\log p_{i|j}$, which is* not *Eq.* (15)*: every pair is weighted identically as a positive, and the positive/negative asymmetry InfoNCE relies on disappears. Single-positive InfoNCE is recovered exactly when π_i is a Dirac distribution on a unique positive (case (a)), not whenever the weights happen to be uniform.*

B PROOF OF THEOREM 1

Proof. Substituting $\log p_{i|j} = \langle \mathbf{z}_i, \mathbf{z}_j \rangle / \tau - \log Z_i$ into Eq. (10) and using $\bar{P}_{ij} = P_{ij}/N$ — which follows from row-stochasticity, since $\sum_{(i,j)\in\widetilde{E}} P_{ij} = \sum_i \big(\sum_j P_{ij}\big) = \sum_i 1 = N$ — gives

$$\mathcal{L}_{\mathrm{OT}} = -\frac{1}{N\tau} \sum_{i,j} P_{ij} \langle \mathbf{z}_i, \mathbf{z}_j \rangle + \frac{1}{2N} \sum_{i,j} P_{ij} (\log Z_i + \log Z_j).$$

The first term is $\mathcal{L}_{\mathrm{align}}^{\mathrm{OT}}$. For the second, split the sum:

$$\sum_{i,j} P_{ij} \log Z_i = \sum_i \log Z_i \big(\sum_j P_{ij} \big) = \sum_i \log Z_i$$

by row-stochasticity, and

$$\sum_{i,j} P_{ij} \log Z_j = \sum_j \log Z_j \big(\sum_i P_{ij} \big) = \sum_j \log Z_j$$

by column-stochasticity (Lemma 1). Both sums equal $\sum_i \log Z_i$, so the second term is

$$\frac{1}{2N} \cdot 2 \sum_i \log Z_i = \mathcal{L}_{\mathrm{unif}}.$$

□