LEARNING SELF-EXPLAINABLE EDGE REPRESENTATIONS VIA COMPOSITIONAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning is central to graph machine learning, with link prediction as a central task. However, most graph models represent edges through opaque similarity functions, offering limited insight into how the latent memberships of two nodes combine to produce, or inhibit, a particular interaction. Many networks naturally admit a compositional view, where nodes express soft membership over multiple latent archetypes and edges may arise through different combinations of these memberships. Motivated by this structure, we propose a compositional graph model that learns self-explainable edge representations. Nodes are represented as simplex-valued compositions over learned archetypes, with each archetype defining an edge channel that combines the shared membership of a node pair in that archetype with their Aitchison agreement over the remaining archetypes. The resulting channel values jointly determine the edge probability and, after normalization, form a compositional representation of the edge itself. This yields explanations by construction: the same quantities used for prediction describe how an interaction is allocated across latent archetypes. We establish that these channels retain relational information lost by global membership overlap and full compositional distance, respond predictably to membership perturbations, and support exact counterfactual reasoning through channel restriction. Across link-prediction benchmarks, the model achieves competitive performance against strong graph-learning baselines while enabling new analyses of how latent memberships shape individual interactions. In particular, its compositional structure supports semantic membership interventions and counterfactual studies that probe which latent channels carry individual edge predictions.

1 INTRODUCTION

Many real-world systems can naturally be represented as graphs, including social networks, communication systems, knowledge networks, and molecular interactions (Hamilton, 2020; Hamilton et al., 2018). Machine learning methods for graph-structured data seek to exploit these relational structures to learn informative representations of entities and their interactions (Chami et al., 2022). A central direction in graph representation learning has therefore been to learn expressive representations that preserve information about both node attributes and structural context (Zhou et al., 2022). More recently, this objective has been extended toward self-explainable representations, in which the learned representation not only supports prediction but also provides an interpretable account of the underlying interaction (Nakis et al., 2026).

A plethora of prior work on graph representation learning has focused on constructing node embeddings that map structural proximity in a graph to geometric proximity in the latent space. Early methods such as DeepWalk (Perozzi et al., 2014) and node2vec (Grover & Leskovec, 2016a) learn representations from random walks over graph neighborhoods, while subsequent work has explored alternative architectures, objectives, and latent geometries (Abu-El-Haija et al., 2018; Duong et al., 2023; Nickel & Kiela, 2017). An important family of approaches instead represents nodes through latent group memberships or structural roles, including stochastic block models and their mixed-membership extensions (Holland et al., 1983; Airoldi et al., 2007), overlapping affiliation models (Yang & Leskovec, 2013), and edge-partition formulations (Zhou, 2015). These models associate graph connectivity with latent structure, but generally do not provide a component-wise account of how individual latent factors contribute to a particular edge prediction. More recent work

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.

addresses link explanation directly by identifying influential neighborhoods, substructures, or connecting paths (Zhu et al., 2023b; Zhang et al., 2023). Such explanations characterize the structural evidence supporting a prediction, rather than decomposing the prediction itself into latent interaction components. Recent work (Nakis et al., 2026), introduced self-explainable compositional node representations via Aitchison distance and geometry (Aitchison, 1982). Here, we extend this perspective to the interaction itself by constructing a self-explainable edge representation whose archetype-specific components directly determine the prediction and provide an intrinsic decomposition of the edge likelihood.

Compositional geometry provides a natural basis for a self-explainable edge mechanism. Each node is represented by a simplex-valued composition whose components encode relative memberships in learned archetypes. Since these memberships sum to a constant, their information is inherently relative, and ratios between components are more meaningful than absolute magnitudes. Aitchison geometry captures this structure through log-ratio comparisons (Aitchison, 1982; Egozcue et al., 2003). AICoG (Nakis et al., 2026) uses this geometry to learn compositional node representations and predict edges from their overall Aitchison distance. We instead model the interaction associated with each archetype separately. For archetype k, shared membership measures how strongly both nodes participate in k, while agreement among the ratios of their remaining memberships captures the compositional context of that shared participation. These archetype-specific channels enter the edge likelihood directly, giving each an explicit contribution to the prediction. The same mechanisms that produce an edge prediction therefore also explain it, moving self-explainability from individual node representations to their interaction.

We introduce the Archetypal Compositional Edge model (ACE), in which each learned archetype defines an interaction channel combining the endpoints' shared membership in that archetype with their Aitchison agreement over the complementary composition. The channels sum to the edge odds, while their normalized values form a compositional representation of the edge itself. Because individual archetypes enter the likelihood explicitly, ACE reduces the rotational ambiguity of distance-based latent representations to permutation of archetype labels. Our contributions are fourfold. **(i)** We introduce an edge representation whose archetype-specific components directly determine the likelihood and show that complementary channels retain interaction information not determined by shared membership and full Aitchison distance alone. **(ii)** We characterize targeted membership interventions, proving that increasing membership in an archetype increases its channel allocation with explicit bounds. **(iii)** We show that the channel allocation exactly determines the effect of removing any subset of channels, including a smallest removal that reduces a prediction below a prescribed threshold. **(iv)** Across graph benchmarks, ACE remains competitive in link prediction while recovering more stable archetypes on channel-generated graphs. On real networks, its edge allocations are sparse and semantically structured, its channel-removal explanations identify the training evidence that most affects predictions after refitting, and membership interventions alter both edge explanations and new-link predictions in the direction predicted by the theory.

2 Related Work

Graph representation learning has largely focused on node embeddings that preserve structural information through spectral, random-walk, and proximity-based objectives (Belkin & Niyogi, 2001; Perozzi et al., 2014; Grover & Leskovec, 2016a; Tang et al., 2015), or through neighborhood aggregation in graph neural networks (Kipf, 2016; Hamilton et al., 2017; Velickovic et al., 2017). For link prediction, endpoint representations are commonly combined through pairwise features or local structural information (Grover & Leskovec, 2016a; Zhang & Chen, 2018; Zhang et al., 2022). A complementary line models connectivity directly through latent memberships, including stochastic block models, mixed-membership and overlapping affiliation models, and edge-partition models (Holland et al., 1983; Airoldi et al., 2007; Yang & Leskovec, 2013; Zhou, 2015). Latent distance models instead relate connectivity to geometric proximity (Hoff et al., 2002), while HM-LDM combines simplex-constrained memberships with latent geometry (Nakis et al., 2022). ACE combines membership and geometry at the edge level, with each archetype defining an interaction that depends on shared membership and geometric agreement among the remaining memberships. Compositional data are naturally represented through log-ratio geometry (Aitchison, 1982; Egozcue et al., 2003), which has been used for nonlinear prediction and component influence (Huang et al., 2023), association-network inference (Friedman & Alm, 2012; Kurtz et al., 2015), and graph-aware

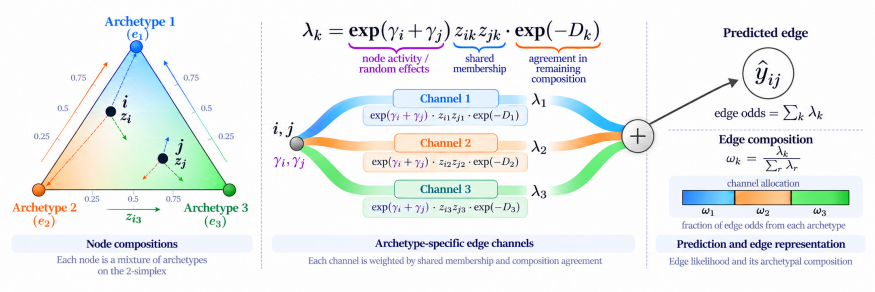


Figure 1: **Overview of ACE.** Each node has a composition over K learned archetypes. For pair (i, j), channel k combines node activity, shared membership $z_{ik}z_{jk}$, and Aitchison agreement over the complementary subcompositions obtained by removing k and reclosing. The channels sum to the edge odds, while their normalized strengths $\omega_{ij,k} = \lambda_{ij,k} / \sum_r \lambda_{ij,r}$ form the compositional edge representation.

distances and convolution (Kalodikis & Matz, 2024; Lu et al., 2025). Most closely related, AICoG learns compositional node representations and predicts links from their overall Aitchison distance (Nakis et al., 2026). ACE instead resolves this interaction across archetype-specific channels whose contributions enter the edge likelihood directly. Graph explanation methods typically identify supporting subgraphs, paths, neighbors, or prototypes (Ying et al., 2019; Luo et al., 2020; Zhang et al., 2023; Chang et al., 2024; Zhu et al., 2023b;a). ACE places the explanation in the latent edge representation itself, so the same channels that determine the prediction also define its decomposition and support exact interventions on the fitted model, unlike graph-edit counterfactual methods such as CF-GNNExplainer (Lucic et al., 2022).

3 Proposed Method

Preliminaries. Let $\mathcal{G} = (V, E)$ be a simple undirected graph with node set $V = [N]$ and adjacency matrix $\boldsymbol{Y} \in \{0,1\}^{N\times N}$, where $\boldsymbol{Y}_{ij} = \boldsymbol{Y}_{ji}$ and $\boldsymbol{Y}_{ii} = 0$. For any positive integer r, let $[r] = \{1, \dots, r\}$, and define the set of unordered node pairs $\mathcal{P} = \{(i,j) : 1 \le i < j \le N\}$. For $(i,j) \in \mathcal{P}$, $Y_{ij} = (\boldsymbol{Y})_{ij}$ denotes the corresponding edge indicator. Bold uppercase and lowercase symbols denote matrices and vectors, respectively, while non-bold symbols denote scalars. All vectors are columns, and $\mathbf{1}_r$ and $\mathbf{0}_r$ denote the r-dimensional vectors of ones and zeros. Logarithms and exponentials act componentwise, and all logarithms are natural. We adopt an interaction-level view of interpretability. Each node is represented by a composition over K learned archetypes, and channel k combines the endpoints' shared membership in archetype k with agreement between their complementary subcompositions. The channel values jointly determine the edge probability, while their normalized values, conditional on an edge, define its allocation across archetypes. Because the same channels also determine the effect of channel removal, this allocation forms a self-explainable edge representation intrinsic to the prediction.

Compositional Nodes. For a positive integer r and fixed $K \ge 3$, the simplex and closure are

$$\mathcal{S}^{r-1} = \{\boldsymbol{z} \in \mathbb{R}^r_{>0} : \mathbf{1}_r^\top \boldsymbol{z} = 1\}, \qquad \mathcal{C}(\boldsymbol{v}) = \frac{\boldsymbol{v}}{\sum_{k=1}^r v_k}, \quad \boldsymbol{v} \in \mathbb{R}^r_{>0}.$$

Node i has a composition $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iK})^\top \in \mathcal{S}^{K-1}$, where z_{ik} is its relative membership in archetype k. These compositions form the rows of the node-membership matrix

$$\boldsymbol{Z} = [\boldsymbol{z}_1 \ \cdots \ \boldsymbol{z}_N]^\top \in \mathbb{R}^{N\times K}_{>0}, \qquad Z_{ik} = z_{ik}, \quad \boldsymbol{Z}\mathbf{1}_K = \mathbf{1}_N. \tag{1}$$

Each node therefore has one K-part composition with intrinsic dimension $D := K - 1$, or equivalently $K = D + 1$. We use K for both the number of archetypes and edge channels. Node i also has an activity $\gamma_i \in \mathbb{R}$, collected in $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_N)^\top \in \mathbb{R}^N$.

Aitchison geometry compares membership ratios (Aitchison, 1982). The centred log-ratio transform (CLR) is $\mathrm{clr}(\boldsymbol{z}) = \log \boldsymbol{z} - (\mathbf{1}_r^\top \log \boldsymbol{z})\mathbf{1}_r/r$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{r-1}$, the Aitchison distance satisfies

$$d_A(\boldsymbol{x}, \boldsymbol{y})^2 = \|\mathrm{clr}(\boldsymbol{x}) - \mathrm{clr}(\boldsymbol{y})\|_2^2 = \frac{1}{r} \sum_{1\le k<\ell\le r} \left(\log \frac{x_k}{x_\ell} - \log \frac{y_k}{y_\ell}\right)^2, \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm. Thus distance measures disagreement between the endpoints' relative profiles. For node i, the CLR coordinates are $\boldsymbol{c}_i = \mathrm{clr}(\boldsymbol{z}_i) \in \mathbb{R}^K$. These vectors form the rows of the matrix $\boldsymbol{C} = [\boldsymbol{c}_1 \ \cdots \ \boldsymbol{c}_N]^\top \in \mathbb{R}^{N\times K}, \quad C_{ik} = c_{ik}, \quad \boldsymbol{C}\mathbf{1}_K = \mathbf{0}_N$. Each vector satisfies $\mathbf{1}_K^\top \boldsymbol{c}_i = 0$ and $\boldsymbol{z}_i = \mathrm{softmax}(\boldsymbol{c}_i)$, where $\mathrm{softmax}(\boldsymbol{v})_k = e^{v_k}/\sum_{\ell=1}^r e^{v_\ell}$ for $\boldsymbol{v} \in \mathbb{R}^r$.

The CLR transform is a bijective isometry onto $H_r = \{\boldsymbol{c} \in \mathbb{R}^r : \mathbf{1}_r^\top \boldsymbol{c} = 0\}$, which for $r = K$ has intrinsic dimension $D = K - 1$. Alternative log-ratio representations include the isometric log-ratio (ILR) transform, which uses D orthonormal coordinates in $\mathbb{R}^D$ (Egozcue et al., 2003), and the additive log-ratio (ALR) transform, which depends on a chosen reference component (Greenacre et al., 2023). We use CLR because it preserves the Aitchison geometry while retaining one coordinate per archetype, aligning directly with the K edge channels (see also Appendix C).

We denote the overall parameter vector as $\boldsymbol{\theta} = \left(\mathrm{vec}(\boldsymbol{C})^\top, \boldsymbol{\gamma}^\top\right)^\top \in \mathbb{R}^{NK+N}, \quad \boldsymbol{C}\mathbf{1}_K = \mathbf{0}_N$.

Channels as Edge Mechanisms. Channel k combines shared membership in archetype k with agreement among the ratios of the remaining archetypes. For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{K-1}$, $\mathbf{x}_{-k}$ denotes $\mathbf{x}$ with part k removed. The complement distance and channel value are

$$D_k(\mathbf{x}, \mathbf{y}) = d_A(\mathcal{C}(\mathbf{x}_{-k}), \mathcal{C}(\mathbf{y}_{-k})), \qquad M_k(\mathbf{x}, \mathbf{y}) = x_k y_k e^{-D_k(\mathbf{x},\mathbf{y})}. \tag{3}$$

The product $x_k y_k$ measures shared membership, while e^{-D_k} measures agreement between the complementary subcompositions. Each complement has $K - 1 = D$ parts and intrinsic dimension $D - 1$. Removing k and reclosing preserves every ratio among the retained parts, so D_k depends only on how the endpoints compare across the other archetypes. The two factors therefore capture distinct aspects of the interaction. For a node pair, the distance is $D_{ij,k} = D_k(\mathbf{z}_i, \mathbf{z}_j)$, yielding $\lambda_{ij,k} = e^{\gamma_i+\gamma_j} M_k(\mathbf{z}_i, \mathbf{z}_j) = e^{\gamma_i+\gamma_j} z_{ik} z_{jk} e^{-D_{ij,k}}$, $k \in [K]$. The activity factor scales all channels equally and therefore affects only their absolute scale. Because the graph is undirected, the channel is symmetric in its endpoints. The channel also admits a positive-semidefinite kernel interpretation, with shared membership acting as a rank-one kernel and the complement term as a Laplacian kernel in Aitchison coordinates. The CLR difference is $\mathbf{b}_{ij} = \mathbf{c}_i - \mathbf{c}_j$, with kth coordinate $b_{ij,k}$. The complement distances have the closed form $D^2_{ij,k} = \|\mathbf{b}_{ij}\|_2^2 - \frac{K}{K-1} b^2_{ij,k} \geq 0$. Thus, all K complement distances can be computed in $O(K)$ time per pair. For a fixed pair, a larger $|b_{ij,k}|$ means that removing archetype k removes more of the compositional disagreement, resulting in a smaller complement distance. This ordering concerns only the complement term. The full channel also depends on the shared membership $z_{ik} z_{jk}$ (see also Appendix C).

Likelihood and allocation. The channel values form the vector $\boldsymbol{\lambda}_{ij} = (\lambda_{ij,1}, \ldots, \lambda_{ij,K})^\top$, with total odds $\Lambda_{ij} = \mathbf{1}_K^\top \boldsymbol{\lambda}_{ij}$. The categorical variable $T_{ij} \in \{0, 1, \ldots, K\}$ encodes the edge assignment, where zero denotes no edge and $k > 0$ an edge assigned to channel k. Assignments at distinct pairs are conditionally independent, with $\mathbb{P}(T_{ij} = 0) = \frac{1}{1+\Lambda_{ij}}$, and $\mathbb{P}(T_{ij} = k) = \frac{\lambda_{ij,k}}{1+\Lambda_{ij}}, \quad k \in [K]$. The observed variable is $Y_{ij} = \mathbf{1}\{T_{ij} > 0\}$. Marginalising the assignment gives

$$p_{ij} := \mathbb{P}(Y_{ij} = 1) = \frac{\Lambda_{ij}}{1 + \Lambda_{ij}}, \qquad \lambda_{ij,k} = \frac{\mathbb{P}(T_{ij} = k)}{\mathbb{P}(T_{ij} = 0)}. \tag{4}$$

Thus, $\lambda_{ij,k}$ is the odds of channel k relative to no edge and Λ_{ij} is the total edge odds. The edge representation is the composition $\boldsymbol{\omega}(i, j) := (\omega_1(i, j), \ldots, \omega_K(i, j))^\top \in \mathcal{S}^{K-1}$, with

$$\omega_k(i, j) := \mathbb{P}(T_{ij} = k \mid Y_{ij} = 1) = \frac{\lambda_{ij,k}}{\Lambda_{ij}}, \qquad \pi_{ij,k} := \mathbb{P}(T_{ij} = k) = p_{ij}\omega_k(i, j). \tag{5}$$

Since $T_{ij} = k$ with $k > 0$ implies $Y_{ij} = 1$, Bayes' rule gives

$$\mathbb{P}(T_{ij} = k \mid Y_{ij} = 1) = \frac{\mathbb{P}(T_{ij} = k)}{\mathbb{P}(Y_{ij} = 1)} = \frac{\lambda_{ij,k}}{\Lambda_{ij}}.$$

Thus, $\boldsymbol{\omega}_{ij}$ allocates an observed edge across channels, while $\boldsymbol{\pi}_{ij}$ retains the absolute probability scale and satisfies $\mathbf{1}_K^\top \boldsymbol{\pi}_{ij} = p_{ij}$. Each channel can therefore be read through its shared membership $z_{ik} z_{jk}$, complement distance $D_{ij,k}$, allocation $\omega_k(i, j)$, and probability contribution $\pi_{ij,k}$. Because activity scales all channels equally, it cancels from $\boldsymbol{\omega}_{ij}$ and affects only the absolute quantities p_{ij} and $\boldsymbol{\pi}_{ij}$. The relative allocation between channels k and ℓ satisfies $\log \frac{\omega_k(i,j)}{\omega_\ell(i,j)} = \log \frac{z_{ik} z_{jk}}{z_{i\ell} z_{j\ell}} - (D_{ij,k} -$

$D_{ij,\ell}), \quad k, \ell \in [K]$. Thus, the relative contribution of two channels separates exactly into shared membership and complementary agreement.

Information supplied by the complement. The Hadamard product is a standard dyadic feature (Grover & Leskovec, 2016b). Its closure is $\mathbf{q}_{ij} = \mathcal{C}(\mathbf{z}_i \circ \mathbf{z}_j)$, where $\circ$ denotes componentwise multiplication. A common full-distance penalty in every channel would cancel from equation 5, giving $\boldsymbol{\omega}_{ij} = \mathbf{q}_{ij}$. Complement distances let disagreement change the allocation itself.

Theorem 3.1 (Pairs distinguished by complement channels)**.** *For every $K \geq 3$, there exist $\mathbf{x}, \mathbf{y}, \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}} \in \mathcal{S}^{K-1}$ with identical Hadamard products and identical Aitchison distances, i.e. $\mathbf{x} \circ \mathbf{y} = \widetilde{\mathbf{x}} \circ \widetilde{\mathbf{y}}$, and $d_A(\mathbf{x}, \mathbf{y}) = d_A(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$, but $\boldsymbol{\omega}(\mathbf{x}, \mathbf{y}) \neq \boldsymbol{\omega}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$.*

The theorem establishes that the explanation retains information about which parts account for the endpoint difference after shared membership and total distance have been fixed (proof in Appendix D). For $K \geq 3$, the channel sum $F(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} M_k(\mathbf{x}, \mathbf{y})$, its log-odds $\log F$, and its induced probability $F/(1+F)$ admit no exact finite-dimensional bilinear representation over the simplex, even with arbitrary nonlinear node transformations and additive node-specific terms. Thus, fixing the number of explanation channels does not restrict ACE to a finite-dimensional bilinear interaction (proof and details in Appendix E).

Response to membership interventions. An interpretable edge representation should respond predictably when a node's association with a particular archetype is deliberately changed. We therefore consider a targeted compositional intervention that increases the relative membership of one endpoint in archetype k while preserving the ratios among all remaining archetypes. Because this perturbation also changes competing complement distances, it is not immediate that channel k should gain explanatory mass. The following result shows that it necessarily does.

The endpoints $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{K-1}$, channel $k \in [K]$, and activities are fixed. With $\mathbf{e}_k$ denoting the kth standard basis vector in $\mathbb{R}^K$, a perturbation of the first endpoint is

$$\mathbf{x}(t) = \mathcal{C}(\mathbf{x} \circ \exp(t\mathbf{e}_k)), \qquad t \in \mathbb{R}. \tag{6}$$

This multiplies x_k/x_ℓ by e^t for every $\ell \neq k$ and preserves the ratios among the remaining parts. The resulting allocation is $\omega_k(t)$, obtained from equation 5 at endpoints $\mathbf{x}(t), \mathbf{y}$. The bound below uses $\kappa_K = \sqrt{(K-2)/(K-1)}$ and the allocation log-odds $L_k(t) = \log \frac{\omega_k(t)}{1-\omega_k(t)}$.

Theorem 3.2 (Monotone channel allocation)**.** *For the channel allocation defined by equation 3 and for every $t_1, t_2 \in \mathbb{R}$ with $t_2 > t_1$,*

$$(1-\kappa_K)(t_2 - t_1) \leq L_k(t_2) - L_k(t_1) \leq (1+\kappa_K)(t_2 - t_1). \tag{7}$$

In particular, $\omega_k(t)$ is strictly increasing in t.

Theorem 3.2 gives the edge representation a directionally consistent response to targeted membership interventions: increasing relative membership in archetype k necessarily increases the share of the explanation assigned to channel k, even though the competing complement distances also change. This provides a controlled way to probe how changes in a node's latent composition reshape its interactions. The guarantee concerns the relative channel allocation; the total edge probability may still increase or decrease because all channels contribute to it. The bounds also quantify how large an intervention is required to obtain a desired explanatory allocation. For a target $\omega_k(0) < w_\star < 1$, the required increase in allocation log-odds is $\Delta_\star = \log \frac{w_\star}{1-w_\star} - L_k(0) > 0$. There is a unique $t_\star > 0$ satisfying $\omega_k(t_\star) = w_\star$, and

$$\frac{\Delta_\star}{1+\kappa_K} \leq t_\star \leq \frac{\Delta_\star}{1-\kappa_K}. \tag{8}$$

Hence the theorem not only fixes the direction of the response, but also bounds the perturbation needed to attain a prescribed channel allocation. The corresponding Aitchison change is $d_A(\mathbf{x}(t_\star), \mathbf{x}) = t_\star \sqrt{(K-1)/K}$. It is the smallest change attaining the target along the specified one-archetype perturbation path (see Appendix F for the proof).

Explanations by Channel Removal. Channel removal provides an intervention on the fitted edge mechanism: it asks which archetype-specific channels support a prediction and how much the prediction would change if selected channels were disabled. For $S \subseteq [K]$, removal sets $\lambda_{ij,k}$ to zero for $k \in S$ while holding every other channel value fixed, yielding the probability $p_{ij}^{(-S)}$. The original

K-part node compositions and $\boldsymbol{\theta}$ remain fixed. This leads to two natural questions. Given a budget of r removals, which r channels produce the largest decrease in edge probability? Conversely, given a target probability $p_\star$, what is the smallest set of channels whose removal reduces the prediction below that target? The following result answers both questions exactly from the fitted channel allocation.

Theorem 3.3 (Exact and optimal channel removal)**.** *Fix a node pair $(i, j) \in \mathcal{P}$. For $S \subseteq [K]$, define $W_{ij}(S) = \sum_{k\in S} \omega_k(i, j)$. Then,*

$$p_{ij}^{(-S)} = \frac{p_{ij}(1 - W_{ij}(S))}{1 - p_{ij}W_{ij}(S)}, \qquad p_{ij} - p_{ij}^{(-S)} = \frac{p_{ij}(1 - p_{ij})W_{ij}(S)}{1 - p_{ij}W_{ij}(S)}. \tag{9}$$

Moreover, let $\omega_{(1)}(i, j) \geq \cdots \geq \omega_{(K)}(i, j)$ denote the channel allocations in nonincreasing order, with ties ordered arbitrarily. For every $r \in \{0, \ldots, K\}$, among all subsets $S \subseteq [K]$ with $|S| = r$, the decrease $p_{ij} - p_{ij}^{(-S)}$ is maximized by removing channels corresponding to $\omega_{(1)}(i, j), \ldots, \omega_{(r)}(i, j)$.

Finally, for any target $0 < p_\star < p_{ij}$, the minimum number of channels whose removal yields $p_{ij}^{(-S)} \leq p_\star$ is

$$r_\star = \min\{r \in [K] : \sum_{k=1}^{r} \omega_{(k)}(i, j) \geq \frac{p_{ij} - p_\star}{p_{ij}(1 - p_\star)}\}. \tag{10}$$

Theorem 3.3 gives the channel allocation an operational interpretation. For a fixed removal budget, the largest allocations identify the channels whose removal causes the greatest decrease in the fitted edge probability. For a target $p_\star$, the ordered prefix satisfying equation 10 gives a smallest channel set that reduces the prediction to that level. Thus, $(p_{ij}, \boldsymbol{\omega}_{ij})$ determines both the effect of any channel removal and a minimal counterfactual explanation, requiring only $O(K \log K)$ sorting.

For a proper subset S, the retained channels have conditional allocation $\omega_k^{(-S)} = \omega_k/(1 - W_S)$ for $k \notin S$. Removing all channels gives probability zero, for which a conditional allocation given an edge is undefined. We provide the proof and the rule for partial attenuation in Appendix G. The probability contribution π_k and the effect of removing channel k are not the same quantity. Removing a single channel changes the edge probability by $(1 - p)\pi_k/(1 - \pi_k)$ because the remaining channels are renormalized. This intervention sets the selected channel to zero while leaving the learned node compositions fixed. It therefore differs from removing an archetype from the compositions, which would also change the complement distances of the remaining channels. We interpret channel removal as a counterfactual of the fitted model rather than a causal intervention on the observed graph. The channel allocation also has a local differential interpretation: $\frac{\partial \operatorname{logit}(p)}{\partial \log \lambda_k} = \omega_k$, and $\frac{\partial p}{\partial \log \lambda_k} = p(1 - p)\omega_k$. Thus, ω_k is also the first-order influence of channel k on the edge log-odds, complementing the finite channel-removal interpretation above. The derivation and corresponding sensitivity to complement distances are given in Appendix G.1.

Identifiability of the archetypes. Latent-distance models are invariant to rotations of the latent coordinates: if predictions depend only on $\|\mathbf{c}_i - \mathbf{c}_j\|_2$, then $\mathbf{c}_i \mapsto \mathbf{Q}\mathbf{c}_i$ leaves the likelihood unchanged. This is harmless for prediction, but problematic for archetype-level explanations, since a rotation mixes the coordinates and changes what "archetype k" represents. ACE removes this rotational symmetry because both $z_{ik}z_{jk}$ and $D_{ij,k}^2 = \|\mathbf{b}_{ij}\|_2^2 - \frac{K}{K-1}b_{ij,k}^2$ refer explicitly to coordinate k; a general rotation therefore changes the channel values and the likelihood. The remaining ambiguity is only permutation of archetype labels, as in mixed-membership models. ACE thereby combines the predictive geometry of latent-distance modeling with the archetype-level identifiability of mixed-membership structure, while additionally providing an explicit channel decomposition of each edge prediction.

Training and complexity. Let $\mathcal{E}_{\text{tr}} \subset \mathcal{P}$ denote the m observed training edges and let $\mathcal{E}_0 = \mathcal{P} \setminus \mathcal{E}_{\text{tr}}$ denote the $U = \binom{N}{2} - m$ training non-edges. We define an unbiased case-control scalable estimator of the loss. Each update uses all m training edges together with a multiset $\mathcal{B}$ of B non-edges drawn independently and uniformly from $\mathcal{E}_0$. We optimize an unbiased case-control estimator using all training edges and uniformly sampled non-edges,

$$\widehat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{B}) = -\frac{1}{m}\left[\sum_{(i,j)\in\mathcal{E}_{\text{tr}}} \log p_{ij} + \frac{U}{B}\sum_{(i,j)\in\mathcal{B}} \log(1 - p_{ij})\right] + \frac{1}{m}\left(\frac{\|\boldsymbol{C}\|_F^2}{2s_c^2} + \frac{\|\boldsymbol{g}\|_2^2}{2s_g^2} + \frac{w_\beta(\beta - \mu_\beta)^2}{2s_\beta^2}\right),$$

Table 2: Link prediction, AUC-ROC (mean of five runs). Bold: best; underline: second best.

	AstroPh				*GrQc*				*HepTh*				*Cora*				*DBLP*			
Dimension (D)	8	16	32	64	8	16	32	64	8	16	32	64	8	16	32	64	8	16	32	64
NODE2VEC	.943	.954	.961	.962	.928	.932	.937	.936	.879	.882	.888	.892	.761	.760	.766	.777	.920	.923	.931	.941
ROLE2VEC	.957	.969	.970	.965	.927	.936	.934	.934	.897	.907	.902	.895	.769	.767	.759	.752	.940	.952	.943	.944
NETMF	.904	.928	.946	.955	.835	.882	.882	.883	.778	.797	.802	.793	.698	.675	.674	.654	.791	.817	.829	.842
SLIM-RAA	.965	.970	.969	.969	.940	.943	.947	.949	.902	.914	.919	.920	.797	.782	.784	.789	.953	.959	.961	.962
MMSBM	.886	.897	.907	.913	.813	.814	.806	.817	.773	.756	.751	.757	.667	.679	.667	.659	.779	.776	.780	.777
DC-MMSBM	.918	.906	.904	.896	.883	.910	.909	.892	.809	.813	.799	.794	.792	.810	.816	.792	.816	.785	.760	.753
MNMF	.877	.916	.938	.954	.881	.905	.918	.918	.810	.844	.864	.875	.694	.718	.700	.680	.859	.898	.919	.931
HM-LDM	.940	.941	.944	.950	.937	.939	.940	.944	.876	.876	.882	.886	.808	.809	.800	.806	.890	.876	.891	.921
AICOG	.965	.972	.975	.976	**.953**	**.958**	**.961**	.961	.911	.927	.928	.929	**.837**	**.846**	**.850**	**.851**	.949	.958	.961	.962
ACE	**.975**	**.982**	**.986**	**.987**	**.953**	**.958**	.960	**.962**	**.920**	**.928**	**.933**	**.935**	.825	.826	.835	.834	**.962**	**.967**	**.969**	**.970**

where we introduce regularization on both the compositional coordinates and node activities via $\bar{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i,\ \boldsymbol{g} = \boldsymbol{\gamma} - \bar{\gamma}\mathbf{1}_N,\ \beta = 2\bar{\gamma}$ (see also Appendix H). The resulting cost is $O((N+m+B)K)$ per update and $O((N+m)K)$ for $B = O(m)$.

4 RESULTS & DISCUSSION

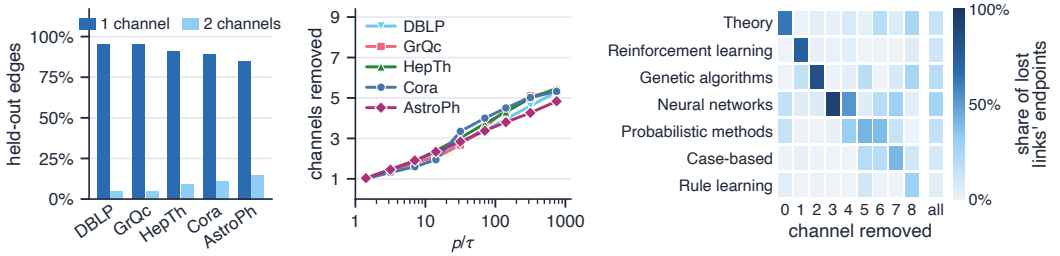
We evaluate ACE on five networks following Nakis et al. (2026): the citation network *Cora* (Sen et al., 2008), and the coauthorship networks *DBLP*, *AstroPh*, *GrQc*, and *HepTh* (Perozzi et al., 2017; Leskovec et al., 2007). The baselines span Euclidean embeddings (NODE2VEC (Grover & Leskovec, 2016a), ROLE2VEC (Ahmed et al., 2018)), matrix factorisation (NETMF (Qiu et al., 2018)), mixed-membership models (MMSBM (Airoldi et al., 2007), MNMF (Wang et al., 2017)) and simplex-based latent distance models (SLIM-RAA (Nakis et al., 2023a), HM-LDM (Nakis et al., 2022), AICOG (Nakis et al., 2026)). We add a degree-corrected MMSBM (DC-MMSBM) with node activities and ACE's logistic link as the block-model comparison for the explanation experiments. All methods share splits and representation sizes $D \in \{8, 16, 32, 64\}$, with $K = D + 1$ for simplex-based models (details in Appendix A). We omit GNN baselines as our setting is featureless and unsupervised, where message passing underperforms without auxiliary features (Nikolentzos et al., 2024). For completeness, we also report node-classification results in Appendix B.

Recovery and identifiability. We evaluate whether the learned archetypes can be recovered consistently by generating graphs from the ACE channel likelihood and the AICOG distance likelihood, then fitting ACE, AICOG, and DC-MMSBM (Table 1). On channel-generated graphs, ACE achieves the lowest membership error (0.51 versus 0.79 and 0.98) and the highest agreement across restarts (BNMI 0.72 versus 0.61 and 0.23). After orthogonal alignment, the error of AICOG drops to 0.51, while that of ACE changes only from 0.51 to 0.46, consistent with the rotational ambiguity of the distance likelihood and the permutation-only ambiguity of ACE. On distance-generated graphs, the two models recover the planted structure similarly after alignment, 0.50 and 0.51.

Table 1: Synthetic recovery ($N = 5000$, $K = 4$, mean degree 16; mean±sd over five graphs). Aligned ℓ_1 uses the best orthogonal map. Restart denotes BNMI across restarts. Bold marks the best result (see also Appendix I).

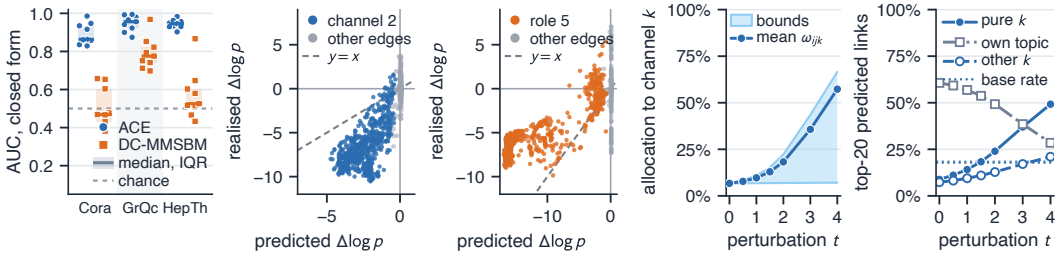
Generator	Model	$\ell_1 \downarrow$	Aligned $\ell_1 \downarrow$	BNMI $\uparrow$	Restart $\uparrow$
Channel	ACE	**.507**±.036	**.459**±.008	**.607**±.020	**.723**±.039
	AICoG	.785±.203	.512±.009	.494±.061	.609±.022
	DC-MMSBM	.983±.038	.928±.031	.300±.027	.229±.043
Distance	ACE	.613±.090	**.497**±.007	.536±.024	**.661**±.015
	AICoG	.786±.152	.509±.003	.470±.044	.590±.018
	DC-MMSBM	1.004±.071	.938±.068	.248±.028	.191±.048

Link prediction. Following Perozzi et al. (2014); Nakis et al. (2023b), we remove 50% of edges while preserving connectivity and pair them with an equal number of sampled non-edges. Models are trained on the residual graph and evaluated over five runs for $D \in \{8, 16, 32, 64\}$. Embedding baselines use the standard binary operators of Grover & Leskovec (2016a) with an L_2-regularized logistic classifier, while likelihood-based models score links directly. Table 2 reports mean AUC-ROC, with standard errors (on the order of 10^{-3}) omitted for clarity. ACE ranks first or second in every setting, with the best or tied-best performance on *AstroPh*, *GrQc*, *HepTh*, and *DBLP* in all but one configuration, while AICOG performs best on *Cora*. Strong performance is retained with the superior identifiability of ACE, whose channel likelihood removes rotational ambiguity and leaves only permutation of archetype labels.



(a) halving p **(b)** falling below τ **(c)** topics of lost links

Figure 2: Channel removal ($K = 9$). (a) Held-out edges whose probability is halved by removing one or two channels. (b) Channels required to reduce a prediction below τ, by confidence p/τ. (c) Ground-truth topics of Cora links lost after removing each channel.



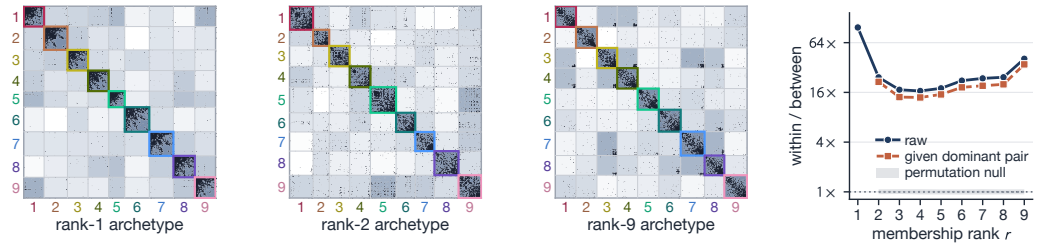
(a) closed form vs. refit **(b)** ACE, channel 2 **(c)** DC-MMSBM, role 5 **(d)** existing links **(e)** predicted links

Figure 3: **Counterfactuals against refitting and membership interventions.** (a) AUC for identifying the 10% of held-out edges most affected by refitting, with one point per channel or role and median and quartiles. (b,c) Predicted versus realised changes on GrQc for the median-AUC ACE channel and DC-MMSBM role. (d) Existing-link allocations on Cora with Theorem 3.2 bounds. (e) Targeted-topic shares among the top-20 predicted new links for topic-pure and other archetypes, with base rates and the node's original-topic share.

Edge-mechanism ablation. We compare four mechanisms within the same training pipeline: shared membership $z_{ik}z_{jk}$, AICOG's distance e^{-d_A}, shared membership times the full distance, and ACE (Appendix J). The three mechanisms containing a distance term differ by at most 0.009 AUC on real graphs, while shared membership alone is 0.02–0.07 lower (Table 8). On *GrQc* and *HepTh*, the difference between ACE and AICOG narrows when the prior is removed (Table 10), indicating that their predictive performance is largely determined by estimation rather than the channel rule. The latent representations differ more substantially. On graphs generated by the ACE likelihood, the error of a distance-only fit drops from 0.78 to 0.47 after orthogonal alignment, while ACE changes only from 0.47 to 0.45 (Table 11). This effect depends on the generating mechanism, as ACE is also rotationally ambiguous on graphs generated from shared membership times the full distance (0.73 to 0.34 after alignment). On real graphs, the complement term reallocates a median of 1–6% of the edge composition relative to the Hadamard closure and changes the dominant channel on 5–11% of held-out edges (Table 12).

Sparse and semantically structured explanations. For held-out edges from the best $K = 9$ fit, using the precision–recall break-even threshold τ, ACE uses on average 1.7–2.2 effective channels per edge. The median dominant allocation is 0.82–0.92, compared with 0.64–0.66 for degree-matched null pairs (Table 12). This concentration also yields short counterfactual explanations, with the minimal removal set given exactly by equation 10 (Figure 2a–b). Removing a single channel halves the probability of 85–95% of held-out edges, while higher-confidence predictions require more channels to be removed. At $p/\tau = 500$, about five of nine channels are needed. The learned channels also exhibit semantic structure. On *Cora*, fitted without topic labels, the links lost after removing a channel are concentrated in a single topic, which accounts for 30–92% of their endpoints compared with 2–23% among the remaining links (Figure 2c, Table 15). Across $K = 9$, 17, and 33, each channel is dominated by one within-topic pair at a median 3.5–3.7 times its base rate (Table 16).

Counterfactuals against refitting & Membership interventions. We test whether the channels identified by ACE correspond to training evidence that affects the same predictions after refitting. For each channel on *Cora*, *GrQc*, and *HepTh*, we remove its attributed training edges, refit ACE three



(a) $r = 1$ (dominant) **(b)** $r = 2$ (second-largest) **(c)** $r = K = 9$ (smallest) **(d)** Connectivity across ranks

Figure 4: **GrQc Relational structure and membership ranks.** ACE with $K = 9$. (a–c) Adjacency reordered by the archetype occupying rank r, nodes grouped by archetype, sorted by membership magnitude. Each block pair is shaded by its edge density (log scale). (d) Within- versus between-archetype connectivity across ranks. The conditioned curve controls for the dominant-archetype pair, and the shaded band is the 95% null interval via dominant-archetype group permutations. Lower-rank enrichment persists after this control.

times from scratch, and compare with removing the same number of random training edges. Held-out edges attributed to the removed channel lose a median 4.5–6.9 in log-probability, compared with 0.02–0.06 for other held-out edges and about 0.5 under random removal. The closed-form removal effect identifies the 10% of held-out edges most affected by refitting with median AUC 0.86–0.96 across channels (Figure 3a), although it underestimates the magnitude of the change (2.0–2.1 versus 4.5–6.9). For DC-MMSBM, the corresponding AUC is 0.47–0.78, and only 29–53% of the most affected edges are assigned to the removed role, compared with 74–84% for ACE. Thus, the channel allocation predicts which held-out edges depend most strongly on the training evidence associated with that channel. We test Theorem 3.2 on labelled *Cora* by increasing one archetype membership for 300 nodes, using a disjoint half of the nodes to map archetypes to topics. Across all 2,201 interventions, the targeted archetype's allocation on existing links increases within the bounds of equation 7. Equation 8 also brackets the required perturbation for all 3,406 evaluated edges (Figure 3d, Table 18). The interventions also redirect link predictions. For topic-pure archetypes, the targeted topic rises from 8.9% to 49.2% among the node's 20 most probable new links, compared with an 18.4% base rate, while the node's original topic falls from 61% to 29% (Figure 3e). AICOG and DC-MMSBM also redirect predictions, but only ACE provides a per-edge explanation that changes consistently with the intervention. At matched membership $z_{ik} = 0.8$, the targeted archetype dominates 91% of the node's existing links, against 3% for DC-MMSBM (Table 18, Appendix N).

Relational structure across membership ranks. We fit ACE with $K = 9$ on *GrQc* and let $k_i^{(r)}$ denote the archetype occupying the r-th largest component of $\mathbf{z}_i$. Figure 4(a–c) reorders the adjacency by $k_i^{(r)}$ for the dominant ($r = 1$), second-largest ($r = 2$), and smallest ($r = K$) memberships. Clear structure remains beyond the dominant partition. We quantify connectivity enrichment at each rank by $\eta_r = P(A_{ij} = 1 \mid k_i^{(r)} = k_j^{(r)}) / P(A_{ij} = 1 \mid k_i^{(r)} \neq k_j^{(r)})$. To account for associations with dominant membership, we also compute an adjusted enrichment η_r^{adj} by stratifying pairs according to their unordered dominant-archetype pair $\{k_i^{(1)}, k_j^{(1)}\}$ and using a Mantel–Haenszel adjusted risk ratio (Greenland & Robins, 1985). For the lower ranks, we construct a permutation null by shuffling rank-r labels within dominant-archetype groups, preserving their conditional distribution while removing residual associations with connectivity. Figure 4d shows enrichment across all ranks. Adjustment produces a modest reduction, and lower-rank enrichment remains above the null. Thus, the learned compositions retain relational structure not explained solely by dominant membership.

5 Conclusion & Limitations

We introduced ACE, a compositional graph model whose archetype-specific channels determine both edge predictions and their explanations. Combining shared membership with complementary Aitchison agreement, the resulting edge allocation supports exact channel removal and bounded responses to membership interventions. ACE remains competitive in link prediction while recovering stable archetypes and producing sparse, semantically structured explanations. Channel removal also identifies predictions most affected by removing attributed training edges and refitting. Our experiments are transductive, although an encoder can extend the model to unseen nodes. Channel removal is a counterfactual of the fitted model rather than a causal intervention on the observed graph.

AI USE STATEMENT

In this work, we used generative AI tools for feedback on the experimental methodology, for assistance with writing code, for drafting proof sketches, and for drafting and editing parts of the paper. We did not use them to formulate the mathematical claims, to develop the conceptual framework or to interpret the results.

We have reviewed all AI-assisted work. All code has been verified by the authors. Furthermore, every proof was derived and checked in full by the authors. Lastly, all experimental design decisions and all interpretation of the results are the authors'. We take full responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors have read the ICLR Code of Ethics and adhere to it. All used datasets are standard in the graph representation learning literature and publicly available. No new data were collected and no human subjects were involved.

REPRODUCIBILITY STATEMENT

The code for the model and its training is included as supplementary material and all datasets used are publicly available. Furthermore, the training setup, hyperparameters, splits and model-selection rule are reported in Appendix A, and the protocol of each experiment in the appendix section that reports it.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alex Alemi. Watch Your Step: Learning Node Embeddings via Graph Attention. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9198–9208, 2018.

Nesreen K Ahmed, Ryan Rossi, John Boaz Lee, Theodore L Willke, Rong Zhou, Xiangnan Kong, and Hoda Eldardiry. Learning role-based graph embeddings. *arXiv preprint arXiv:1802.02896*, 2018.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels, 2007. URL https://arxiv.org/abs/0705.4485.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345821.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

Ines Chami, Sami Abu-El-Haija, Bryan Perozzi, Christopher Ré, and Kevin Murphy. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *Journal of Machine Learning Research*, 23(89):1–64, 2022.

Heng Chang, Jiangnan Ye, Alejo Lopez-Avila, Jinhua Du, and Jia Li. Path-based explanation for knowledge graph completion. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 231–242. ACM, August 2024. doi: 10.1145/3637528.3671683. URL http://dx.doi.org/10.1145/3637528.3671683.

Chi Thang Duong, Thanh Tam Nguyen, Trung-Dung Hoang, Hongzhi Yin, Matthias Weidlich, and Quoc Viet Hung Nguyen. Deep MinCut: Learning Node Embeddings by Detecting Communities. *Pattern Recognition*, 134:109126, 2023.

Juan Jose Egozcue, Vera Pawlowsky-Glahn, Glòria Figueras, and Carles Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300, 04 2003. doi: 10.1023/A:1023818214614.

Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):1–11, 09 2012. doi: 10.1371/journal.pcbi.1002687. URL https://doi.org/10.1371/journal.pcbi.1002687.

Michael Greenacre, Eric Grunsky, John Bacon-Shone, Ionas Erb, and Thomas Quinn. Aitchison's compositional data analysis 40 years on: A reappraisal, 2023. URL https://arxiv.org/abs/2201.05197.

Sander Greenland and James M. Robins. Estimation of a common effect parameter from sparse follow-up data. *Biometrics*, 41 1:55–68, 1985. URL https://api.semanticscholar.org/CorpusID:9979814.

Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016a.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016b. URL https://arxiv.org/abs/1607.00653.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2020.

William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications, 2018. URL https://arxiv.org/abs/1709.05584.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. doi: 10.1198/016214502388618906. URL https://doi.org/10.1198/016214502388618906.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(83)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873383900217.

Shimeng Huang, Elisabeth Ailer, Niki Kilbertus, and Niklas Pfister. Supervised learning and model analysis with compositional data. *PLOS Computational Biology*, 19(6):e1011240, 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011240. URL http://dx.doi.org/10.1371/journal.pcbi.1011240.

Dimitrios Kalodikis and Gerald Matz. Graph signal processing for compositional data. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pp. 413–417, 2024. doi: 10.1109/IEEECONF60004.2024.10942632.

TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004226. URL http://dx.doi.org/10.1371/journal.pcbi.1004226.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

Shan Lu, Huiwen Wang, and Jichang Zhao. Graph convolutional network for compositional data. *Information Fusion*, 117:102798, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102798. URL https://www.sciencedirect.com/science/article/pii/S1566253524005761.

Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks, 2022. URL https://arxiv.org/abs/2102.03322.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network, 2020. URL https://arxiv.org/abs/2011.04573.

Nikolaos Nakis, Abdulkadir Çelikkanat, and Morten Mørup. Hm-ldm: A hybrid-membership latent distance model, 2022. URL https://arxiv.org/abs/2206.03463.

Nikolaos Nakis, Abdulkadir Celikkanat, Louis Boucherie, Christian Djurhuus, Felix Burmester, Daniel Mathias Holmelund, Monika Frolcová, and Morten Mørup. Characterizing polarization in social networks using the signed relational latent distance model. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11489–11505. PMLR, 25–27 Apr 2023a. URL https://proceedings.mlr.press/v206/nakis23a.html.

Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann Jørgensen, and Morten Mørup. A hierarchical block distance model for ultra low-dimensional graph representations, 2023b. URL https://arxiv.org/abs/2204.05885.

Nikolaos Nakis, Chrysoula Kosma, Panagiotis Promponas, Michail Chatzianastasis, and Giannis Nikolentzos. Aitchison embeddings for learning compositional graph representations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Maximilian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6341–6350, 2017.

Giannis Nikolentzos, Michail Chatzianastasis, and Michalis Vazirgiannis. What do gnns actually learn? towards understanding their representations, 2024. URL https://arxiv.org/abs/2304.10851.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014.

Bryan Perozzi, Vivek Kulkarni, Haochen Chen, and Steven Skiena. Don't walk, skip! online learning of multi-scale network embeddings. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 258–265, 2017.

Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, pp. 459–467, 2018.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 2008.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pp. 1067–1077, 2015.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pp. 587–596, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318693. doi: 10.1145/2433396.2433471. URL https://doi.org/10.1145/2433396.2433471.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks, 2018. URL https://arxiv.org/abs/1802.09691.

Muhan Zhang, Pan Li, Yinglong Xia, Kai Wang, and Long Jin. Labeling trick: A theory of using graph neural networks for multi-node representation learning, 2022. URL https://arxiv.org/abs/2010.16103.

Shichang Zhang, Jiani Zhang, Xiang Song, Soji Adeshina, Da Zheng, Christos Faloutsos, and Yizhou Sun. Page-link: Path-based graph neural network explanation for heterogeneous link prediction, 2023. URL https://arxiv.org/abs/2302.12465.

Jingya Zhou, Ling Liu, Wenqi Wei, and Jianxi Fan. Network Representation Learning: From Preprocessing, Feature Extraction to Node Embedding. *ACM Computing Surveys*, 55(2):1–35, 2022.

Mingyuan Zhou. Infinite edge partition models for overlapping community detection and link prediction, 2015. URL https://arxiv.org/abs/1501.06218.

Huaisheng Zhu, Dongsheng Luo, Xianfeng Tang, Junjie Xu, Hui Liu, and Suhang Wang. Self-explainable graph neural networks for link prediction, 2023a. URL https://arxiv.org/abs/2305.12578.

Huaisheng Zhu, Dongsheng Luo, Xianfeng Tang, Junjie Xu, Hui Liu, and Suhang Wang. Self-explainable graph neural networks for link prediction, 2023b. URL https://arxiv.org/abs/2305.12578.

A Experimental Details

Table 3: Dataset statistics. N: nodes; $|E|$: edges; C: classes ("–": unlabelled).

	LastFM	*Citeseer*	*Cora*	*DBLP*	*AstroPh*	*GrQc*	*HepTh*
N	7,624	3,327	2,708	27,199	17,903	5,242	8,638
$\|E\|$	55,612	9,104	5,278	66,832	197,031	14,496	24,827
C	18	6	7	–	–	–	–

For link prediction we hold out half of the edges, drawn uniformly among the edges outside a random spanning tree such that the training graph stays connected, and sample an equivalent amount of non-edges uniformly. All methods share the same splits. For the labelled-graph explanation analyses the models are fitted on the full graph without using the labels.

ACE. Node coordinates and activities are optimised jointly with Adam at learning rate 0.01 for 5,000 epochs, with the rate quartered after 4,000. During the first 500 epochs only the activity scale is fitted. Each update uses all training edges and five times as many non-edges, drawn uniformly with replacement, with the factor U/B of equation 38 making the sampled term unbiased. Sampled pairs that are training edges are discarded from the non-edge term, except for AICoG, which samples uniformly over all pairs. Held-out pairs are never excluded for any model. Thus as in Nakis et al. (2026), an edge removed by the split may be sampled as a non-edge, so the model treats the held-out half of the graph as unobserved zeros rather than conditioning on it. In all experiments $s_g = s_\gamma$, $\mu_\beta = 2\mu_\gamma$ and $w_\beta/s_\beta^2 = N/(4s_\gamma^2)$, such that the Gaussian penalty of equation 37 reduces to $\|\mathbf{C}\|_F^2/(2s_c^2) + \|\boldsymbol{\gamma} - \mu_\gamma \mathbf{1}_N\|_2^2/(2s_\gamma^2)$, with $s_c = 2$ on the compositional coordinates and $s_\gamma = 3$ on the activities around $\mu_\gamma = -3$. These are fixed for every graph and every K, and the mechanism ablation of Appendix J removes the prior by setting both scales to 100. Coordinates are initialised from a spectral embedding of the training adjacency, except where random initialisation is stated. The complement distance is evaluated as $\sqrt{\max\{D_{ij,k}^2, \varepsilon\}}$ with $\varepsilon = 10^{-12}$. The training objective reported for model selection is the mean loss over the last 200 epochs.

Hardware. All fits run on a single NVIDIA RTX A5000. One fit of 5,000 epochs takes about one minute on GrQc at any K, and between 1 and 4 minutes on AstroPh, the largest graph, going from $K = 9$ to $K = 65$.

B Node Classification

Table 4: Node classification, micro-F1 (mean over runs). Every model is fitted on the full graph without labels; its memberships in centred log-ratio coordinates are fixed features for a multinomial logistic regression and for k-nearest neighbours, on the same 40/10/50 stratified split. Under kNN the distance is the Aitchison distance between the fitted compositions.

	Cora				*Citeseer*				*LastFM*			
Dimension (D)	8	16	32	64	8	16	32	64	8	16	32	64
Logistic regression												
ACE	.767	.783	.798	.811	.653	.656	.681	.722	.849	.856	.863	.865
AICoG	**.773**	**.806**	**.827**	**.831**	**.680**	**.702**	**.719**	**.734**	**.857**	**.867**	**.870**	**.871**
MMSBM	.387	.581	.735	.782	.495	.590	.643	.689	.496	.702	.787	.834
k-nearest neighbours												
ACE	.835	.841	.843	**.842**	**.742**	**.749**	.743	.742	.863	.863	.867	.866
AICoG	**.844**	**.845**	**.845**	.840	.741	.744	**.745**	**.744**	**.867**	**.870**	**.871**	**.872**
MMSBM	.304	.379	.440	.442	.414	.421	.435	.457	.316	.323	.359	.394

Given that classification is not part of the explainability claim, we only report it as evidence for the representation's usefulness. We compare ACE with AICoG, the strongest compositional model in Table 2, and with the MMSBM, the mixed-membership baseline fitted under our training pipeline. Each model is fitted on the full labelled graph without using the labels, and its memberships in centred log-ratio coordinates are fixed features for a multinomial logistic regression and for k-nearest neighbours on the same 40/10/50 stratified split (Table 4). Under kNN the distance is the Aitchison distance between the fitted compositions.

Under logistic regression AICoG performs better than ACE by 0.005–0.046 micro-F1, and both are above the MMSBM. Under kNN, AICoG and ACE are within 0.009 in every cell. Thus, the two representations carry comparable class information, and the gap under the linear classifier reflects how the class information is organised in the representation rather than its absence. The MMSBM demonstrates the opposite pattern. It improves sharply with D under logistic regression from 0.39 to 0.78 on Cora, but remains at 0.30–0.46 under kNN. Furthermore, the validation selects $k = 3$ for MMSBM, while selecting 20 and 50 for ACE and AICoG, respectively. The MMSBM's near-hard memberships place nodes in a few corners of the simplex, and consequently the neighbourhood says little about a node even when a linear boundary can separate the corners. This near-hard structure is also consistent with the instability observed for the degree-corrected variant under refitting (Appendix M).

C Complement Geometry

CLR and ILR coordinates. Recall that K counts composition parts and $D = K-1$ is the intrinsic latent dimension. Let $\mathbf{V} \in \mathbb{R}^{K\times D}$ have orthonormal columns spanning H_K, so $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_D$ and $\mathbf{V}^\top\mathbf{1}_K = \mathbf{0}_D$; $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. The ILR vector $\mathbf{u}_i = \mathbf{V}^\top\mathbf{c}_i \in \mathbb{R}^D$ satisfies $\mathbf{c}_i = \mathbf{V}\mathbf{u}_i$ and $\|\mathbf{u}_i-\mathbf{u}_j\|_2 = \|\mathbf{c}_i-\mathbf{c}_j\|_2$. Thus CLR and ILR represent the same Euclidean geometry. Stacking ILR vectors as rows gives the matrix representation

$$\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_N]^\top \in \mathbb{R}^{N\times D}, \qquad \mathbf{U} = \mathbf{C}\mathbf{V}, \quad \mathbf{C} = \mathbf{U}\mathbf{V}^\top. \tag{11}$$

Each row of $\mathbf{U}$ contains D unconstrained coordinates, whereas the corresponding row of $\mathbf{C}$ contains $K = D + 1$ coordinates with sum zero. A baseline with latent dimension D therefore matches the number of independent compositional coordinates when $K = D + 1$.

Complement identity. Fix a pair (i, j) and write $\mathbf{b} = \mathbf{c}_i - \mathbf{c}_j$, $b_k = b_{ij,k}$, and $D_k = D_{ij,k}$. Because $\sum_{\ell=1}^{K} b_\ell = 0$, the CLR difference between the complementary subcompositions has entries $b_\ell + b_k/(K-1)$ for $\ell \neq k$. Closure changes all log entries by the same constant, which cancels on centring. Expanding the squared norm gives

$$D_k^2 = \sum_{\ell \neq k} \left(b_\ell + \frac{b_k}{K-1} \right)^2 = \|\mathbf{b}\|_2^2 - \frac{K}{K-1} b_k^2. \tag{12}$$

This proves the complement identity. The sum of squares is nonnegative and vanishes exactly when the two complementary subcompositions agree. The restriction $K \geq 3$ ensures that each complement has at least two parts; a one-part complement has zero Aitchison distance.

Ordering and retained information. For a fixed pair, equation 12 gives

$$\arg\max_{k \in [K]} D_k = \arg\min_{k \in [K]} |b_k|, \qquad \arg\min_{k \in [K]} D_k = \arg\max_{k \in [K]} |b_k|. \tag{13}$$

These equalities include all tied indices. Here $|b_k|$ is the difference in centred log membership, so the ordering must be read relative to the endpoints' geometric means. Summing the squared distances and rearranging also gives

$$\sum_{k=1}^{K} D_k^2 = \frac{K(K-2)}{K-1} \|\mathbf{b}\|_2^2, \qquad b_k^2 = \frac{K-1}{K} \left(\|\mathbf{b}\|_2^2 - D_k^2 \right). \tag{14}$$

Consequently, the vector of complement distances determines the magnitude of every CLR difference coordinate, as well as the full distance. Its sum alone retains only the full squared distance. Signs are lost: $\mathbf{b}$ and $-\mathbf{b}$ always give the same complement distances.

Shared and differential coordinates. Let $\mathbf{a} = (\mathbf{c}_i + \mathbf{c}_j)/2$ and $\mathbf{D} = (D_1, \ldots, D_K)^\top \in \mathbb{R}^K_{\geq 0}$; this bold vector collects complement distances, while the scalar $D = K - 1$ denotes latent dimension. The pair $(\mathbf{a}, \mathbf{b})$ determines both endpoints: $\mathbf{c}_i = \mathbf{a} + \mathbf{b}/2$ and $\mathbf{c}_j = \mathbf{a} - \mathbf{b}/2$. The Aitchison midpoint $\mathbf{m}_{ij} = \mathcal{C}((\sqrt{z_{ik} z_{jk}})_{k=1}^K)$ satisfies $\mathrm{clr}(\mathbf{m}_{ij}) = \mathbf{a}$. The shared-membership allocation and channel posterior are

$$\mathbf{q}_{ij} = \mathrm{softmax}(2\mathbf{a}), \qquad \boldsymbol{\omega}_{ij} = \mathrm{softmax}(2\mathbf{a} - \mathbf{D}). \tag{15}$$

The vectors $\mathbf{a}$ and $\mathbf{b}$ can be chosen independently in H_K: each choice gives valid endpoints by softmax. Hence there is no forced opposition between the ranking of shared membership and that of complement compatibility. The two rankings can reinforce or oppose one another. The complement-distance ordering in equation 13 concerns $\mathbf{b}$.

For the absolute prediction, define $\mathrm{LSE}(\mathbf{v}) = \log \sum_k e^{v_k}$. Substitution into the channel sum gives

$$\log \frac{p_{ij}}{1 - p_{ij}} = \gamma_i + \gamma_j + \mathrm{LSE}(2\mathbf{a} - \mathbf{D}) - \mathrm{LSE}(\mathbf{a} + \mathbf{b}/2) - \mathrm{LSE}(\mathbf{a} - \mathbf{b}/2). \tag{16}$$

The endpoint normalisers cancel from the posterior and remain in the absolute edge odds. Exchanging the endpoints replaces $\mathbf{b}$ by $-\mathbf{b}$ and preserves $\mathbf{a}$, $\mathbf{D}$, $\boldsymbol{\lambda}_{ij}$, p_{ij}, and $\boldsymbol{\omega}_{ij}$.

D Proof of the Representation Result

Proof of Theorem 3.1. Let $\mathbf{e}_k \in \mathbb{R}^K$ denote the kth standard basis vector. Fix $t > 0$ and define

$$\mathbf{b} = t(\mathbf{e}_1 - \mathbf{e}_2), \qquad \widetilde{\mathbf{b}} = t(\mathbf{e}_1 - \mathbf{e}_3).$$

Construct two pairs of compositions by $\mathbf{x} = \mathrm{softmax}(\mathbf{b}/2)$, $\mathbf{y} = \mathrm{softmax}(-\mathbf{b}/2)$, and $\widetilde{\mathbf{x}} = \mathrm{softmax}(\widetilde{\mathbf{b}}/2)$, $\widetilde{\mathbf{y}} = \mathrm{softmax}(-\widetilde{\mathbf{b}}/2)$.

All four softmax normalising constants are $a = e^{t/2} + e^{-t/2} + K - 2$. Hence, for every $k \in [K]$, $x_k y_k = \widetilde{x}_k \widetilde{y}_k = a^{-2}$, and therefore

$$\mathbf{x} \circ \mathbf{y} = \widetilde{\mathbf{x}} \circ \widetilde{\mathbf{y}}.$$

For the Aitchison distance, notice that

$$d_A(\mathbf{x}, \mathbf{y}) = \|\mathbf{b}\|_2 = \sqrt{2}\, t = \|\widetilde{\mathbf{b}}\|_2 = d_A(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}).$$

Thus the two pairs have the same Hadamard product and the same Aitchison distance. At the same time, for the first pair,

$$D_1(\mathbf{x}, \mathbf{y}) = D_2(\mathbf{x}, \mathbf{y}) = t\sqrt{\frac{K-2}{K-1}},$$

and $D_k(\mathbf{x}, \mathbf{y}) = \sqrt{2}\, t,\ k \geq 3$. For the second pair,

$$D_1(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = D_3(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = t\sqrt{\frac{K-2}{K-1}},$$

and $D_k(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \sqrt{2}\, t,\ k \notin \{1, 3\}$.

Next, by equation 40 and equation 5, for arbitrary compositions $\mathbf{u}, \mathbf{v} \in \mathcal{S}^{K-1}$,

$$\omega_k(\mathbf{u}, \mathbf{v}) = \frac{u_k v_k e^{-D_k(\mathbf{u}, \mathbf{v})}}{\sum_{\ell=1}^{K} u_\ell v_\ell e^{-D_\ell(\mathbf{u}, \mathbf{v})}}.$$

In our construction, $x_k y_k = \widetilde{x}_k \widetilde{y}_k = a^{-2}$ for every k, so the common factor a^{-2} cancels from numerator and denominator. Moreover, if we denote

$$d_- = t\sqrt{\frac{K-2}{K-1}}, \qquad \text{and} \qquad d_+ = \sqrt{2}\, t.$$

Then, both allocations, $\boldsymbol{\omega}(x, y)$ and $\boldsymbol{\omega}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$, have the same denominator,

$$Z = 2e^{-d_-} + (K-2)e^{-d_+}.$$

Therefore,

$$\omega_2(\mathbf{x}, \mathbf{y}) = \frac{e^{-d_-}}{Z}, \qquad \omega_3(\mathbf{x}, \mathbf{y}) = \frac{e^{-d_+}}{Z},$$

whereas

$$\omega_2(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \frac{e^{-d_+}}{Z}, \qquad \omega_3(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \frac{e^{-d_-}}{Z}.$$

Since $K \geq 3$ and $t > 0$, it follows that $d_- < d_+$. Thus,

$$\omega_2(\mathbf{x}, \mathbf{y}) > \omega_3(\mathbf{x}, \mathbf{y}), \qquad \omega_2(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) < \omega_3(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}),$$

and therefore $\boldsymbol{\omega}(\mathbf{x}, \mathbf{y}) \neq \boldsymbol{\omega}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$, completing the proof.

□

The archetype labels are held fixed in this comparison. The constructed pairs have equal total odds when their summed activities agree. The result concerns the information in the channel allocation; predictive gains require empirical evidence.

E Separation from Finite Bilinear Decoders

We examine whether the interaction defined by the channels can be represented exactly by a bilinear decoder after transforming the node compositions. The channel sum and the corresponding edge probability when the pairwise activity term satisfies $\gamma_i + \gamma_j = 0$ are

$$F(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} M_k(\mathbf{x}, \mathbf{y}), \qquad p_0(\mathbf{x}, \mathbf{y}) = \frac{F(\mathbf{x}, \mathbf{y})}{1 + F(\mathbf{x}, \mathbf{y})}. \tag{17}$$

Here F is the total edge odds with zero summed activity, $\log F$ is the corresponding log-odds, and p_0 is the edge probability. Considering all three quantities distinguishes the interaction function from the choice of probability scale.

Theorem E.1 (No exact finite-dimensional bilinear representation)**.** *For fixed $K \geq 3$ and each $G \in \{F, \log F, p_0\}$, there are no finite integer $d \geq 1$, maps $\boldsymbol{\phi}, \boldsymbol{\psi} : \mathcal{S}^{K-1} \to \mathbb{R}^d$, matrix $\mathbf{B} \in \mathbb{R}^{d \times d}$, and functions $a, b : \mathcal{S}^{K-1} \to \mathbb{R}$ such that*

$$G(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{B} \boldsymbol{\psi}(\mathbf{y}) + a(\mathbf{x}) + b(\mathbf{y}) \tag{18}$$

for every $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{K-1}$.

Importantly, the maps ϕ, ψ and the additive terms a, b may be nonlinear.

The hypothetical decoder dimension d is unrestricted and may differ from the model's latent dimension $D = K - 1$. Thus, fixing the number of explanation channels does not impose a finite-dimensional bilinear interaction. The result also covers bilinear log-odds with node activities, since these contribute additive node-specific terms. The comparison concerns exact equality over the simplex.

Proof of Theorem E.1. For $t \in \mathbb{R}$, define $v(t) = t(\mathbf{e}_1 - \mathbf{e}_2)$ and consider the one-parameter family $\mathbf{r}(t) = \mathrm{softmax}(v(t))$. We first compute its CLR coordinates explicitly. Let $Z(t) = \sum_{\ell=1}^K e^{v_\ell(t)}$. Then

$$r_k(t) = \frac{e^{v_k(t)}}{Z(t)},$$

and therefore $\log r_k(t) = v_k(t) - \log Z(t)$. Since $\sum_{k=1}^K v_k(t) = 0$, we have

$$\frac{1}{K} \sum_{\ell=1}^K \log r_\ell(t) = \frac{1}{K} \sum_{\ell=1}^K v_\ell(t) - \log Z(t) = -\log Z(t).$$

Hence, for every $k \in [K]$,

$$\mathrm{clr}(\mathbf{r}(t))_k = \log r_k(t) - \frac{1}{K} \sum_{\ell=1}^K \log r_\ell(t) = v_k(t).$$

Thus, $\mathrm{clr}(\mathbf{r}(t)) = t(\mathbf{e}_1 - \mathbf{e}_2)$. It follows that, for $s, t \in \mathbb{R}$,

$$\mathbf{b} := \mathrm{clr}(\mathbf{r}(s)) - \mathrm{clr}(\mathbf{r}(t)) = (s - t)(\mathbf{e}_1 - \mathbf{e}_2).$$

Consequently, $\|\mathbf{b}\|_2^2 = 2(s-t)^2$, while $b_1 = s - t,\ b_2 = -(s-t),\ b_k = 0 \quad (k \geq 3)$. Using

$$D_k(\mathbf{r}(s), \mathbf{r}(t))^2 = \|\mathbf{b}\|_2^2 - \frac{K}{K-1} b_k^2.$$

Therefore,

$$D_1(\mathbf{r}(s), \mathbf{r}(t))^2 = D_2(\mathbf{r}(s), \mathbf{r}(t))^2 = \left(2 - \frac{K}{K-1}\right)(s-t)^2 = \frac{K-2}{K-1}(s-t)^2,$$

whereas, for $k \geq 3$, $D_k(\mathbf{r}(s), \mathbf{r}(t))^2 = 2(s-t)^2$. Thus,

$$D_k(\mathbf{r}(s), \mathbf{r}(t)) = \rho_k |s - t|, \qquad \rho_1 = \rho_2 = \sqrt{\frac{K-2}{K-1}}, \qquad \rho_k = \sqrt{2}, \quad k \geq 3. \tag{19}$$

Since $K \geq 3$, every ρ_k is strictly positive. Fix $s \in \mathbb{R}$, and define

$$h_s(t) = F(\mathbf{r}(s), \mathbf{r}(t)) = \sum_{k=1}^K r_k(s) r_k(t) e^{-\rho_k |s-t|}. \tag{20}$$

Each $r_k(t)$ is smooth on $\mathbb{R}$, and $e^{-\rho_k|s-t|}$ is smooth on $(-\infty, s)$ and (s, ∞). Hence h_s is smooth away from $t = s$. For each $k \in [K]$, set

$$H_{k,s}(t) = r_k(s) r_k(t) e^{-\rho_k |s-t|},$$

so that $h_s(t) = \sum_{k=1}^K H_{k,s}(t)$. For $t > s$,

$$H'_{k,s}(t) = r_k(s) e^{-\rho_k (t-s)} \left(r'_k(t) - \rho_k r_k(t)\right).$$

Taking $t \downarrow s$ gives

$$H'_{k,s}(s^+) = r_k(s)r'_k(s) - \rho_k r_k(s)^2. \tag{21}$$

Similarly, for $t < s$,

$$H'_{k,s}(t) = r_k(s)e^{-\rho_k(s-t)}\left(r'_k(t) + \rho_k r_k(t)\right).$$

Taking $t \uparrow s$ gives

$$H'_{k,s}(s^-) = r_k(s)r'_k(s) + \rho_k r_k(s)^2. \tag{22}$$

Subtracting equation 22 from equation 21,

$$H'_{k,s}(s^+) - H'_{k,s}(s^-) = -2\rho_k r_k(s)^2.$$

Since the sum in equation 20 is finite,

$$h'_s(s^+) - h'_s(s^-) = -2\sum_{k=1}^{K} \rho_k r_k(s)^2 < 0. \tag{23}$$

The strict inequality follows because $\rho_k > 0$ and $r_k(s) > 0$ for every k. Now let

$$g(u) \in \left\{u, \log u, \frac{u}{1+u}\right\}.$$

These choices correspond respectively to $G = F,\ G = \log F,\ G = p_0$. Since every summand in equation 20 is positive, $h_s(t) > 0$ for every $s, t \in \mathbb{R}$. Moreover, $g'(u) > 0$ for every $u > 0$ for each of the three choices of g.

The function h_s is continuous at $t = s$, and its two one-sided derivatives exist. Applying the one-sided chain rule gives

$$(g \circ h_s)'(s^+) = g'(h_s(s))h'_s(s^+) \qquad \text{and} \qquad (g \circ h_s)'(s^-) = g'(h_s(s))h'_s(s^-).$$

Therefore, by equation 23,

$$(g \circ h_s)'(s^+) - (g \circ h_s)'(s^-) = g'(h_s(s))\left(h'_s(s^+) - h'_s(s^-)\right) = -2g'(h_s(s))\sum_{k=1}^{K} \rho_k r_k(s)^2 < 0. \tag{24}$$

Thus, for every fixed s, the function $t \mapsto G(\mathbf{r}(s), \mathbf{r}(t))$ is smooth on $\mathbb{R} \setminus \{s\}$ and has distinct left and right derivatives at $t = s$. Choose distinct points $s_1, \ldots, s_n \in \mathbb{R}$.

We claim that the functions $t \mapsto G(\mathbf{r}(s_i), \mathbf{r}(t)),\ i = 1, \ldots, n$, are linearly independent. Suppose that

$$\sum_{i=1}^{n} c_i G(\mathbf{r}(s_i), \mathbf{r}(t)) = 0 \qquad \text{for every } t \in \mathbb{R}. \tag{25}$$

Fix $j \in \{1, \ldots, n\}$. Because the points $s_1, \ldots, s_n$ are distinct, for every $i \neq j$ the function $t \mapsto G(\mathbf{r}(s_i), \mathbf{r}(t))$ is smooth in a neighbourhood of $t = s_j$. Hence its right and left derivatives at s_j are equal. Taking the right derivative and the left derivative of equation 25 at $t = s_j$, and subtracting, therefore yields

$$c_j \left[\frac{d}{dt}G(\mathbf{r}(s_j), \mathbf{r}(t))\bigg|_{t=s_j^+} - \frac{d}{dt}G(\mathbf{r}(s_j), \mathbf{r}(t))\bigg|_{t=s_j^-}\right] = 0.$$

By equation 24, the quantity in brackets is strictly negative. Thus $c_j = 0$. Since j was arbitrary, $c_1 = \cdots = c_n = 0$, which proves the claimed linear independence.

Suppose now, toward a contradiction, that equation 18 holds for some finite d. Substituting $\mathbf{x} = \mathbf{r}(s),\ \mathbf{y} = \mathbf{r}(t)$, gives

$$G(\mathbf{r}(s), \mathbf{r}(t)) = \boldsymbol{\phi}(\mathbf{r}(s))^\top \mathbf{B}\boldsymbol{\psi}(\mathbf{r}(t)) + a(\mathbf{r}(s)) + b(\mathbf{r}(t)).$$

For fixed s, write $\mathbf{u}(s) = \mathbf{B}^\top \boldsymbol{\phi}(\mathbf{r}(s)) \in \mathbb{R}^d$. The

$$\boldsymbol{\phi}(\mathbf{r}(s))^\top \mathbf{B}\boldsymbol{\psi}(\mathbf{r}(t)) = \sum_{\ell=1}^{d} u_\ell(s)\psi_\ell(\mathbf{r}(t)).$$

Consequently, as a function of t,

$$G(\mathbf{r}(s), \mathbf{r}(t)) \in \operatorname{span}\{\psi_1(\mathbf{r}(t)), \ldots, \psi_d(\mathbf{r}(t)), b(\mathbf{r}(t)), 1\}.$$

This vector space has dimension at most $d + 2$.

On the other hand, the preceding argument shows that for every n and every choice of distinct $s_1, \ldots, s_n$, the functions

$$t \longmapsto G(\mathbf{r}(s_i), \mathbf{r}(t)), \qquad i = 1, \ldots, n,$$

are linearly independent. Choosing $n > d + 2$ gives a contradiction. Therefore, no representation of the form equation 18 exists. □

The theorem concerns the exact unsquared distances in equation 3. Its proof uses their derivative jump along the family in equation 19. Replacing the norm by a smooth approximation with a strictly positive smoothing parameter changes the decoder and requires a separate argument.

F Membership Perturbations and Channel Allocation

Proof of Theorem 3.2. Fix $t_2 > t_1$ and write $\Delta t = t_2 - t_1$. Removing the kth coordinate from $\mathbf{x}(t)$ eliminates the perturbation before closure, so

$$\mathcal{C}(\mathbf{x}(t)_{-k}) = \mathcal{C}(\mathbf{x}_{-k}).$$

Together with equation 3, this gives

$$D_k(\mathbf{x}(t), \mathbf{y}) = d_A(\mathcal{C}(\mathbf{x}_{-k}), \mathcal{C}(\mathbf{y}_{-k})),$$

so $D_k(\mathbf{x}(t), \mathbf{y})$ is constant in t. Now fix $\ell \in [K] \setminus k$ and define

$$\mathbf{c}_\ell(t) = \operatorname{clr}(\mathcal{C}(\mathbf{x}(t)_{-\ell})) - \operatorname{clr}(\mathcal{C}(\mathbf{y}_{-\ell})).$$

Then, $D_\ell(\mathbf{x}(t), \mathbf{y}) = \|\mathbf{c}_\ell(t)\|_2$. We next compute the change in $\mathbf{c}_\ell(t)$ between t_1 and t_2. Write

$$Z(t) = x_k e^t + \sum_{m \neq k} x_m.$$

For every coordinate $r \in [K] \setminus \ell$,

$$\log x_r(t) = \log x_r + t\mathbf{1}\{r = k\} - \log Z(t).$$

Hence,

$$\log x_r(t_2) - \log x_r(t_1) = \Delta t\, \mathbf{1}\{r = k\} - \big(\log Z(t_2) - \log Z(t_1)\big).$$

The average of these increments over the $K - 1$ remaining coordinates is

$$\frac{\Delta t}{K-1} - \big(\log Z(t_2) - \log Z(t_1)\big).$$

Therefore,

$$[\operatorname{clr}(\mathcal{C}(\mathbf{x}(t_2)_{-\ell})) - \operatorname{clr}(\mathcal{C}(\mathbf{x}(t_1)_{-\ell}))]_r = \Delta t \left(\mathbf{1}\{r = k\} - \frac{1}{K-1}\right).$$

Since the $\mathbf{y}_{-\ell}$ term in $\mathbf{c}_\ell(t)$ does not depend on t, it follows that

$$\mathbf{c}_\ell(t_2) - \mathbf{c}_\ell(t_1) = \Delta t\, \mathbf{h}_{\ell,k},$$

where, for $r \in [K] \setminus \ell$, $(h_{\ell,k})_r = \mathbf{1}\{r = k\} - \frac{1}{K-1}$. The vector $\mathbf{h}_{\ell,k}$ has one coordinate equal to $1 - 1/(K-1)$ and $K - 2$ coordinates equal to $-1/(K-1)$. Thus

$$\|\mathbf{h}_{\ell,k}\|_2^2 = \left(1 - \frac{1}{K-1}\right)^2 + (K-2)\left(\frac{1}{K-1}\right)^2 = \frac{K-2}{K-1} = \kappa_K^2. \tag{26}$$

By the reverse triangle inequality,

$$\begin{aligned} |D_\ell(\mathbf{x}(t_2), \mathbf{y}) - D_\ell(\mathbf{x}(t_1), \mathbf{y})| &= |\|\mathbf{c}_\ell(t_2)\|_2 - \|\mathbf{c}_\ell(t_1)\|_2| \\ &\leq \|\mathbf{c}_\ell(t_2) - \mathbf{c}_\ell(t_1)\|_2 \\ &= \kappa_K \Delta t. \end{aligned} \tag{27}$$

For brevity, write $D_\ell(t) = D_\ell(\mathbf{x}(t), \mathbf{y})$. For $\ell \neq k$, the definition of $\mathbf{x}(t)$ gives

$$\frac{x_k(t)}{x_\ell(t)} = e^t \frac{x_k}{x_\ell}.$$

Consequently,

$$\log \frac{x_k(t_2)}{x_\ell(t_2)} - \log \frac{x_k(t_1)}{x_\ell(t_1)} = \Delta t.$$

Using equation 28

$$\log \frac{\omega_k(i,j)}{\omega_{ij,\ell}} = \log \frac{z_{ik} z_{jk}}{z_{i\ell} z_{j\ell}} - (D_{ij,k} - D_{ij,\ell}), \qquad k, \ell \in [K]. \tag{28}$$

and the fact that $D_k(\mathbf{x}(t), \mathbf{y})$ is constant in t, we obtain

$$\log \frac{\omega_k(t_2)}{\omega_\ell(t_2)} - \log \frac{\omega_k(t_1)}{\omega_\ell(t_1)} = \Delta t + D_\ell(t_2) - D_\ell(t_1). \tag{29}$$

Combining this identity with equation 27 gives

$$(1 - \kappa_K)\Delta t \leq \log \frac{\omega_k(t_2)}{\omega_\ell(t_2)} - \log \frac{\omega_k(t_1)}{\omega_\ell(t_1)} \leq (1 + \kappa_K)\Delta t. \tag{30}$$

Define

$$R_k(t) = \sum_{\ell \neq k} \frac{\omega_\ell(t)}{\omega_k(t)} = \frac{1 - \omega_k(t)}{\omega_k(t)}.$$

Rearranging equation 30 and exponentiating gives, for every $\ell \neq k$,

$$e^{-(1+\kappa_K)\Delta t} \frac{\omega_\ell(t_1)}{\omega_k(t_1)} \leq \frac{\omega_\ell(t_2)}{\omega_k(t_2)} \leq e^{-(1-\kappa_K)\Delta t} \frac{\omega_\ell(t_1)}{\omega_k(t_1)}.$$

Summing these inequalities over $\ell \neq k$ yields

$$e^{-(1+\kappa_K)\Delta t} R_k(t_1) \leq R_k(t_2) \leq e^{-(1-\kappa_K)\Delta t} R_k(t_1). \tag{31}$$

Finally,

$$L_k(t) = \log \frac{\omega_k(t)}{1 - \omega_k(t)} = -\log R_k(t).$$

Taking logarithms in equation 31 and multiplying by -1 gives

$$(1 - \kappa_K)\Delta t \leq L_k(t_2) - L_k(t_1) \leq (1 + \kappa_K)\Delta t,$$

which is equation 7. Since

$$\kappa_K = \sqrt{\frac{K-2}{K-1}} < 1,$$

the lower bound is strictly positive whenever $t_2 > t_1$. Thus $L_k(t_2) > L_k(t_1)$. Since

$$\omega_k(t) = \frac{e^{L_k(t)}}{1 + e^{L_k(t)}}$$

and the map $u \mapsto e^u/(1+e^u)$ is strictly increasing, $\omega_k(t_2) > \omega_k(t_1)$. Therefore, $\omega_k(t)$ is strictly increasing in t. □

Target allocation and perturbation size. The channel values are positive and continuous in t, so $L_k(t)$ is continuous. The lower bound in equation 7 implies $L_k(t) \to +\infty$ as $t \to +\infty$. Together with strict monotonicity, this gives a unique $t_\star > 0$ for every target $w_\star \in (\omega_k(0), 1)$. Applying equation 7 with $t_1 = 0$ and $t_2 = t_\star$ yields

$$(1 - \kappa_K)t_\star \leq \Delta_\star \leq (1 + \kappa_K)t_\star,$$

which rearranges to equation 8.

Finally,

$$\mathrm{clr}(\mathbf{x}(t)) - \mathrm{clr}(\mathbf{x}) = t\left(\mathbf{e}_k - \frac{1}{K}\mathbf{1}_K\right).$$

Taking its norm gives $d_A(\mathbf{x}(t), \mathbf{x}) = |t|\sqrt{(K-1)/K}$. Strict monotonicity means that the target is attained exactly for $t \geq t_\star$, proving minimality along the specified perturbation path. No optimisation over other composition changes is asserted.

G ALLOCATION, REMOVAL, AND SENSITIVITY

Probability allocation. Under equation 32

$$\mathbb{P}(T_{ij}=0)=\frac{1}{1+\Lambda_{ij}},\qquad \mathbb{P}(T_{ij}=k)=\frac{\lambda_{ij,k}}{1+\Lambda_{ij}},\quad k\in[K]. \tag{32}$$

the categories $1,\dots,K$ are mutually exclusive and exhaust the event $Y_{ij}=1$. Marginalising, gives $p_{ij}=\Lambda_{ij}/(1+\Lambda_{ij})$ and Bayes' rule gives $\omega_{ij,k}=\lambda_{ij,k}/\Lambda_{ij}$. Consequently, $\pi_{ij,k}=p_{ij}\omega_{ij,k}$ and $\sum_k \pi_{ij,k}=p_{ij}$. The latent assignment is not observed in the binary adjacency. Its posterior is determined by the fitted model.

Proof of Theorem 3.3. Fix $(i,j)\in\mathcal{P}$ and $S\subseteq[K]$. Since $\omega_k(i,j)=\frac{\lambda_{ij,k}}{\Lambda_{ij}}$, we have

$$W_{ij}(S)=\sum_{k\in S}\omega_k(i,j)=\frac{1}{\Lambda_{ij}}\sum_{k\in S}\lambda_{ij,k}.$$

Removing the channels in S therefore leaves total odds $\Lambda_{ij}^{(-S)}=\sum_{k\notin S}\lambda_{ij,k}=\Lambda_{ij}\big(1-W_{ij}(S)\big)$. Hence,

$$p_{ij}^{(-S)}=\frac{\Lambda_{ij}^{(-S)}}{1+\Lambda_{ij}^{(-S)}}=\frac{\Lambda_{ij}(1-W_{ij}(S))}{1+\Lambda_{ij}(1-W_{ij}(S))}.$$

Using $\Lambda_{ij}=\frac{p_{ij}}{1-p_{ij}}$, we obtain

$$p_{ij}^{(-S)}=\frac{p_{ij}(1-W_{ij}(S))}{1-p_{ij}W_{ij}(S)}.$$

Since $0<p_{ij}<1$ and $0\le W_{ij}(S)\le 1$,

$$1-p_{ij}W_{ij}(S)\ge 1-p_{ij}>0.$$

Therefore,

$$p_{ij}-p_{ij}^{(-S)}=p_{ij}-\frac{p_{ij}(1-W_{ij}(S))}{1-p_{ij}W_{ij}(S)}=\frac{p_{ij}(1-p_{ij})W_{ij}(S)}{1-p_{ij}W_{ij}(S)},$$

which proves equation 9.

For fixed p_{ij} now, define

$$\Delta(W)=\frac{p_{ij}(1-p_{ij})W}{1-p_{ij}W},\qquad 0\le W\le 1.$$

Then,

$$\Delta'(W)=\frac{p_{ij}(1-p_{ij})}{(1-p_{ij}W)^2}>0,$$

so the decrease in edge probability is strictly increasing in W. Let $\omega_{(1)}(i,j)\ge\cdots\ge\omega_{(K)}(i,j)$ be the channel allocations in nonincreasing order, as in the statement of the Theorem. For every $S\subseteq[K]$ with $|S|=r$,

$$W_{ij}(S)=\sum_{k\in S}\omega_k(i,j)\le\sum_{k=1}^{r}\omega_{(k)}(i,j).$$

Since $\Delta(W)$ is strictly increasing, the decrease $p_{ij}-p_{ij}^{(-S)}$ is maximized by removing channels corresponding to $\omega_{(1)}(i,j),\dots,\omega_{(r)}(i,j)$.

Finally, let $0<p_\star<p_{ij}$. Since $1-p_{ij}W_{ij}(S)>0$, we have

$$\begin{aligned} p_{ij}^{(-S)}\le p_\star \quad &\Longleftrightarrow \quad \frac{p_{ij}(1-W_{ij}(S))}{1-p_{ij}W_{ij}(S)}\le p_\star \\ &\Longleftrightarrow \quad p_{ij}-p_\star\le p_{ij}(1-p_\star)W_{ij}(S) \\ &\Longleftrightarrow \quad W_{ij}(S)\ge\frac{p_{ij}-p_\star}{p_{ij}(1-p_\star)}. \end{aligned}$$

For every subset $S \subseteq [K]$ with $|S| = r$,

$$W_{ij}(S) \leq \sum_{k=1}^{r} \omega_{(k)}(i,j).$$

Thus a subset of r channels can satisfy $p_{ij}^{(-S)} \leq p_\star$ if and only if

$$\sum_{k=1}^{r} \omega_{(k)}(i,j) \geq \frac{p_{ij} - p_\star}{p_{ij}(1-p_\star)}.$$

The smallest such r is therefore

$$r_\star = \min\left\{ r \in [K] : \sum_{k=1}^{r} \omega_{(k)}(i,j) \geq \frac{p_{ij} - p_\star}{p_{ij}(1-p_\star)} \right\},$$

which proves equation 10. □

Retained allocation and partial attenuation. For a proper subset S, the remaining channel values sum to $\Lambda(1 - W_S) > 0$. Normalising them gives $\omega_k^{(-S)} = \omega_k/(1 - W_S)$ for $k \notin S$. The retained vector is a composition over the surviving channels. When represented in K coordinates, removed entries are zero.

More generally, choose attenuation factors $\mathbf{t} = (t_1, \ldots, t_K)^\top \in [0,1]^K$ and set $\lambda_k^{(\mathbf{t})} = t_k \lambda_k$, keeping the original parameters fixed. Define $\rho = \sum_k t_k \omega_k$. Then

$$p^{(\mathbf{t})} = \frac{p\rho}{1 - p + p\rho}, \qquad \omega_k^{(\mathbf{t})} = \frac{t_k \omega_k}{\rho} \quad (\rho > 0). \tag{33}$$

These formulas follow from $\Lambda^{(\mathbf{t})} = \Lambda\rho$. When $\rho = 0$, the edge probability is zero and a conditional allocation given an edge is undefined. Taking $t_k = 0$ on S and $t_k = 1$ elsewhere recovers channel removal.

For a single removed channel, direct subtraction also gives

$$p - p^{(-\{k\})} = \frac{\lambda_k}{(1+\Lambda)(1+\Lambda-\lambda_k)} = \frac{(1-p)\pi_k}{1-\pi_k}. \tag{34}$$

The allocation π_k is a categorical probability. The removal effect includes the change in the denominator.

G.1 Local Sensitivity

Define $\mathbf{h} = (\log\lambda_1, \ldots, \log\lambda_K)^\top$. Then $\log\{p/(1-p)\} = \mathrm{LSE}(\mathbf{h})$. Differentiating with respect to h_k, holding the other entries fixed, gives $\partial\,\mathrm{LSE}(\mathbf{h})/\partial h_k = \omega_k$. Differentiating the logistic map gives the second identity in equation 35

$$\frac{\partial\ \mathrm{logit}(p)}{\partial \log\lambda_k} = \omega_k, \qquad \frac{\partial p}{\partial \log\lambda_k} = p(1-p)\omega_k. \tag{35}$$

One may also treat the distances as separate inputs to $h_k = \gamma_i + \gamma_j + \log z_{ik} + \log z_{jk} - D_k$. At fixed memberships, activities, and other distance inputs,

$$\frac{\partial \log\{p/(1-p)\}}{\partial D_k} = -\omega_k, \qquad \frac{\partial p}{\partial D_k} = -p(1-p)\omega_k. \tag{36}$$

These are derivatives of the final probability calculation. The full model constrains the distances through the same node compositions, so independently changing a distance need not correspond to a node perturbation. Along any differentiable parameter path, the total derivative is $\mathrm{d}\log\{p/(1-p)\} = \sum_k \omega_k\,\mathrm{d}\log\lambda_k$, which includes the effect on every channel.

H Training Objective and Numerical Evaluation

The training adjacency $\mathbf{A}^{\mathrm{tr}} = (a_{ij}^{\mathrm{tr}}) \in \{0,1\}^{N\times N}$ is symmetric with zero diagonal. Its positive and zero pairs are $\mathcal{E}_{\mathrm{tr}} = \{(i,j) \in \mathcal{P} : a_{ij}^{\mathrm{tr}} = 1\}$ and $\mathcal{E}_0 = \mathcal{P} \setminus \mathcal{E}_{\mathrm{tr}}$, with $m = |\mathcal{E}_{\mathrm{tr}}| > 0$ and $U = |\mathcal{E}_0| = \binom{N}{2} - m > 0$.

Prior. The mean activity is $\bar{\gamma} = N^{-1}\sum_i \gamma_i$, with $\beta = 2\bar{\gamma}$ and $\mathbf{g} = \boldsymbol{\gamma} - \bar{\gamma}\mathbf{1}_N$. Then $\boldsymbol{\gamma} = (\beta/2)\mathbf{1}_N + \mathbf{g}$ and $\mathbf{1}_N^\top \mathbf{g} = 0$. These are an equivalent representation of the activity parameters. We use the Gaussian penalty

$$\mathcal{R}(\boldsymbol{\theta}) = \frac{\|\mathbf{C}\|_F^2}{2s_c^2} + \frac{\|\mathbf{g}\|_2^2}{2s_g^2} + \frac{w_\beta(\beta-\mu_\beta)^2}{2s_\beta^2}, \tag{37}$$

where $\|\mathbf{C}\|_F^2 = \sum_{i=1}^N \sum_{k=1}^K c_{ik}^2$. The scales $s_c, s_g, s_\beta > 0$, weight $w_\beta > 0$, and location $\mu_\beta \in \mathbb{R}$ are fixed hyperparameters. The Gaussian priors on $\mathbf{C}$ and $\mathbf{g}$ are defined on their respective sum-zero subspaces.

Case-control loss. Each update includes all m positive pairs and $B \geq 1$ independent, uniform draws from $\mathcal{E}_0$. The resulting multiset is $\mathcal{B}$, with repeated draws counted separately. The objective minimised during training is

$$\widehat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{B}) = -\frac{1}{m}\left[\sum_{(i,j)\in\mathcal{E}_{\mathrm{tr}}} \log p_{ij} + \frac{U}{B}\sum_{(i,j)\in\mathcal{B}} \log(1-p_{ij})\right] + \frac{\mathcal{R}(\boldsymbol{\theta})}{m}. \tag{38}$$

Here p_{ij} depends on $\boldsymbol{\theta}$ through equation 40 and equation 39

$$p_{ij} := \mathbb{P}(Y_{ij}=1) = \frac{\Lambda_{ij}}{1+\Lambda_{ij}}, \qquad \lambda_{ij,k} = \frac{\mathbb{P}(T_{ij}=k)}{\mathbb{P}(T_{ij}=0)}, \tag{39}$$

$$\lambda_{ij,k} = e^{\gamma_i+\gamma_j} M_k(\mathbf{z}_i,\mathbf{z}_j) = e^{\gamma_i+\gamma_j} z_{ik}z_{jk}e^{-D_{ij,k}}, \qquad k \in [K]. \tag{40}$$

The factor U/B makes the sampled zero term unbiased for its full sum. The prior is included once per update. Node coordinates and activities are optimised together, subject to $\mathbf{C}\mathbf{1}_K = \mathbf{0}_N$; the categorical assignments are marginalised in the likelihood.

Unbiasedness and complexity. The full regularised objective for the specified training adjacency is

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{m}\left[\sum_{(i,j)\in\mathcal{E}_{\mathrm{tr}}} \log p_{ij} + \sum_{(i,j)\in\mathcal{E}_0} \log(1-p_{ij})\right] + \frac{\mathcal{R}(\boldsymbol{\theta})}{m}. \tag{41}$$

Let (I_b, J_b), $b = 1,\dots,B$, be the independent uniform controls. For any real-valued function f on $\mathcal{E}_0$,

$$\mathbb{E}\left[\frac{U}{B}\sum_{b=1}^B f(I_b,J_b)\right] = \frac{U}{B}\sum_{b=1}^B \frac{1}{U}\sum_{(i,j)\in\mathcal{E}_0} f(i,j) = \sum_{(i,j)\in\mathcal{E}_0} f(i,j).$$

Applying this to $f(i,j) = \log(1-p_{ij})$ proves $\mathbb{E}[\widehat{\mathcal{L}}] = \mathcal{L}$ at fixed parameters. At differentiable parameter values the finite sums give the same unbiasedness for the gradients. Given the sampled pairs, computing the likelihood and its gradients costs $O((m+B)K)$. Rowwise softmax and the full prior add $O(NK)$, giving $O((N+m+B)K)$ per update, or $O((N+m)K)$ when B is proportional to m. Using $K = D+1$, the general update cost is equivalently $O((N+m+B)(D+1))$. These bounds concern loss and gradient evaluation; sampling cost depends on its implementation.

Evaluation in log space. For each node, $\log z_{ik} = c_{ik} - \mathrm{LSE}(\mathbf{c}_i)$. Compute $h_{ij,k} = \gamma_i + \gamma_j + \log z_{ik} + \log z_{jk} - D_{ij,k}$ and $\ell_{ij} = \mathrm{LSE}(\mathbf{h}_{ij})$, where $\mathbf{h}_{ij} = (h_{ij,1},\dots,h_{ij,K})^\top$. Then $\log p_{ij} = -\operatorname{softplus}(-\ell_{ij})$ and $\log(1-p_{ij}) = -\operatorname{softplus}(\ell_{ij})$, with $\operatorname{softplus}(x) = \log(1+e^x)$. Stable log-sum-exp and softplus implementations avoid forming large odds or subtracting probabilities close to one.

I SYNTHETIC RECOVERY STUDY

We test whether fitted memberships and channel allocations recover the parameters that generated a graph.

Table 5: Realised synthetic graphs (largest connected component; mean±sd over the five graphs of each setting). Target: expected mean degree set by bisection before the component is taken.

Generator	Target degree	N	$\lvert E\rvert$	Mean degree
Channel	8	4,701±24	20,004±209	8.51±0.09
Channel	16	4,897±6	39,939±288	16.31±0.13
Distance	16	4,865±6	39,861±234	16.39±0.11

Generative process. For each node $i \in [N]$ a dominant archetype is drawn as $d_i \sim \mathrm{Unif}([K])$ and

$$\tilde{\mathbf{z}}_i \sim \mathrm{Dir}\big(\alpha_0 \mathbf{1}_K + \alpha_1 \mathbf{e}_{d_i}\big), \qquad \mathbf{c}_i = s\,\frac{\mathrm{clr}(\tilde{\mathbf{z}}_i)}{\sigma}, \qquad \mathbf{z}_i = \mathrm{softmax}(\mathbf{c}_i), \tag{42}$$

where σ is the root mean square of all CLR coordinates, so that s sets the membership sharpness. Activities are $\gamma_i = \tilde{\gamma}_i + \beta/2$ with $\tilde{\gamma}_i \sim \mathcal{N}(0, \sigma_\gamma^2)$. Edges are drawn independently, $Y_{ij} \sim \mathrm{Bern}(p_{ij})$, from one of four generators:

$$\text{channel:} \quad \lambda_{ij,k} = e^{\gamma_i+\gamma_j} z_{ik} z_{jk} e^{-D_{ij,k}} \text{ (equation 40)}, \tag{43}$$

$$\text{distance:} \quad \mathrm{logit}\, p_{ij} = \gamma_i + \gamma_j - d_A(\mathbf{z}_i, \mathbf{z}_j), \tag{44}$$

$$\text{product:} \quad \lambda_{ij,k} = e^{\gamma_i+\gamma_j} z_{ik} z_{jk}, \tag{45}$$

$$\text{product} \times \text{full:} \quad \lambda_{ij,k} = e^{\gamma_i+\gamma_j} z_{ik} z_{jk} e^{-d_A(\mathbf{z}_i,\mathbf{z}_j)}, \tag{46}$$

with $p_{ij} = \Lambda_{ij}/(1+\Lambda_{ij})$ for the three channel-based generators. The channel generator is ACE's likelihood and the distance generator AICoG's. The offset β is set by bisection so that the expected mean degree over all N nodes equals $\bar{d}$. For every edge, a channel is drawn as $T_{ij} \mid Y_{ij} = 1 \sim \mathrm{Cat}(\boldsymbol{\omega}_{ij})$ under the generator's own posterior, which is the Hadamard closure $\mathbf{q}_{ij}$ for the two product generators; the distance generator has no channels, and its labels, drawn under the channel form, serve only as a reference. We keep the largest connected component, hold out half of the edges, chosen uniformly among the edges outside a random spanning tree so that the training graph stays connected, and sample as many non-edges uniformly within the component. We use $N = 5000$, $K = 4$, $s = 2$, $\alpha_0 = 0.7$, $\alpha_1 = 3$, $\sigma_\gamma = 0.6$ and $\bar{d} \in \{8, 16\}$; Table 5 reports the realised graphs. The main study uses the channel and distance generators, and the mechanism ablation (Appendix J) one graph from each of the four at $\bar{d} = 16$.

Design. In the main study, one set of memberships and activities are drawn and used to generate five graphs per setting, such that the planted truth is identical across graphs, and differences only reflect edge sampling and the split. In the supplementary study, each of the five graphs per setting is generated from independently drawn memberships and activities. ACE, AICoG and DC-MMSBM are fitted to each graph with five random restarts, where the fit with the lowest training objective is kept. Furthermore, ACE is also fitted once per channel graph at $\bar{d} = 16$ started from the planted memberships as a reference.

Metrics. Membership recovery is measured by the mean ℓ_1 distance between fitted and planted compositions after matching archetypes with the Hungarian algorithm on the ℓ_1 cost, and by the Bayesian normalised mutual information (BNMI) between the soft membership matrices, which needs no matching. Channel recovery is evaluated for ACE on the held-out edges of channel graphs, as the top-1 agreement between $\arg\max_k \hat{\omega}_{ij,k}$ and $\arg\max_k \omega_{ij,k}$ after matching, and as the BNMI between the fitted allocations and the drawn channels. Because the draws are noisy, the BNMI of the planted allocations against the same draws serves as a ceiling (oracle).

Identifiability and alignment. A likelihood that depends on the coordinates only through distances, such as AICoG's and the distance generator's, is invariant under every orthogonal map of the CLR coordinates, whereas ACE's likelihood is invariant only under permutations of the archetypes.

Table 6: Synthetic recovery, main study: one set of planted memberships, five graphs per setting ($N = 5000$, $K = 4$), each fit the best of five random restarts by training objective; mean±sd over graphs. Aligned: after the best rotation, reflection and translation of the fitted CLR coordinates onto the planted ones (an upper bound). Top-1 and BNMI draw: channel recovery on held-out edges; parentheses: oracle BNMI of the planted allocations. Bold as in Table 1.

Generator	Degree	Model	ROC ↑	ℓ_1 ↓	ℓ_1 aligned ↓	BNMI ↑	Restart agreement ↑	Top-1 ↑	BNMI draw (oracle) ↑
Channel	8	ACE	**.757**±.010	**.845**±.011	**.714**±.035	**.271**±.029	**.265**±.041	.497±.014	.172±.021 (.659)
		AICoG	.730±.010	1.036±.057	.874±.035	.173±.018	.102±.015	–	–
		DC-MMSBM	.627±.004	1.275±.035	1.225±.060	.027±.014	.010±.002	–	–
Channel	16	ACE	.853±.002	**.507**±.036	**.459**±.008	**.607**±.020	**.723**±.039	.762±.037	.419±.030 (.661)
		AICoG	.850±.001	.785±.203	.512±.009	.494±.061	.609±.022	–	–
		DC-MMSBM	.782±.006	.983±.038	.928±.031	.300±.027	.229±.043	–	–
		ACE, planted start	.857±.002	.427±.010	.426±.010	.672±.015	–	.833±.008	.489±.011 (.661)
Distance	16	ACE	.838±.002	.613±.090	**.497**±.007	.536±.024	**.661**±.015	–	–
		AICoG	.840±.001	.786±.152	.509±.003	.470±.044	.590±.018	–	–
		DC-MMSBM	.775±.003	1.004±.071	.938±.068	.248±.028	.191±.048	–	–

Table 7: Synthetic recovery, supplementary study: as Table 6, but every graph is generated from independently drawn memberships and activities.

Generator	Degree	Model	ROC ↑	ℓ_1 ↓	ℓ_1 aligned ↓	BNMI ↑	Restart agreement ↑	Top-1 ↑	BNMI draw (oracle) ↑
Channel	8	ACE	**.757**±.008	**.813**±.038	**.709**±.022	**.277**±.023	**.248**±.051	.519±.051	.187±.021 (.650)
		AICoG	.717±.009	1.013±.069	.916±.031	.156±.023	.083±.016	–	–
Channel	16	ACE	.851±.004	**.505**±.023	**.464**±.011	**.603**±.020	**.689**±.021	.764±.025	.415±.026 (.653)
		AICoG	.847±.003	.857±.084	.522±.005	.469±.021	.625±.026	–	–
		ACE, planted start	.855±.003	.433±.008	.432±.008	.661±.012	–	.828±.007	.478±.012 (.653)
Distance	16	ACE	.836±.003	.662±.089	.503±.012	.508±.034	**.665**±.013	–	–
		AICoG	.839±.001	.764±.101	.513±.007	.463±.024	.571±.056	–	–

The matched ℓ_1 and BNMI therefore measure recovery including orientation. To separate orientation from the configuration itself, we also map the fitted CLR coordinates onto the planted ones by the best rotation, reflection and translation in the sum-zero subspace (orthogonal Procrustes) and recompute both measures. The map uses the truth, so aligned values are an upper bound on what any post-hoc alignment could achieve. On distance graphs the planted orientation cannot be identified from data by any model, so only aligned values are compared there. Restart agreement is the mean pairwise BNMI between the memberships of the five restarts on a graph; it compares fits with each other and needs no truth.

Results. Tables 6 and 7 report the full results, and Table 5 the realised graphs. At $\bar{d} = 8$, ACE recovers the planted memberships more closely than AICoG before and after alignment (ℓ_1 0.85 against 1.04, aligned 0.71 against 0.87), although recovery is well below that at $\bar{d} = 16$ (BNMI 0.27 against 0.61) and restarts agree less (0.27 against 0.72). AICoG's restarts differ mainly by orientation. On held-out edges, ACE's dominant channel matches the planted one on 76% of edges (83% from the planted start), and its BNMI with the drawn channels is 0.42 against an oracle of 0.66. The DC-MMSBM, fitted by maximum likelihood without a prior, is furthest from the truth, and its restarts disagree (0.01–0.23) as its near-hard memberships settle in different partitions; at $\bar{d} = 8$ it also predicts held-out edges poorly (ROC 0.63). The supplementary study reproduces every ordering between ACE and AICoG. Individual entries differ by at most 0.07, the largest being AICoG's unaligned ℓ_1 at $\bar{d} = 16$, which also varies most across graphs.

J MECHANISM ABLATION

All mechanisms share activities, prior, initialisation and training. Table 8 uses five spectral runs per cell, as Table 2; the prior comparison and the synthetic study use 25 random-initialisation seeds per cell, grouped into five repetitions of five, each keeping the seed with the lowest training objective.

Table 8: Mechanism ablation on real graphs ($D = 8$, AUC-ROC, mean of five runs). All mechanisms share activities, prior, initialisation and training.

Mechanism	AstroPh	GrQc	HepTh	Cora	DBLP
$z_{ik}z_{jk}$	.936	.895	.868	.805	.893
e^{-d_A}	.975	.959	.929	.833	.966
$z_{ik}z_{jk}\,e^{-d_A}$	.976	.955	.924	.825	.964
$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.975	.953	.920	.825	.962

Table 9: ACE against AICoG's distance mechanism fitted within ACE's pipeline (AUC-ROC, mean of five runs).

	AstroPh				*GrQc*				*HepTh*				*Cora*				*DBLP*			
Dimension (D)	8	16	32	64	8	16	32	64	8	16	32	64	8	16	32	64	8	16	32	64
ACE	.975	.982	.986	.987	.953	.958	.960	.962	.920	.928	.933	.935	.825	.826	.835	.834	.962	.967	.969	.970
Distance only	.975	.982	.986	.987	.959	.962	.963	.964	.929	.937	.938	.939	.833	.838	.836	.834	.966	.970	.971	.971

Table 10: Mechanism ablation with and without the Gaussian prior ($D = 8$, AUC-ROC). Each entry: mean$\pm$sd over five repetitions, each the best of five random restarts by training objective.

	GrQc		HepTh	
Mechanism	prior	no prior	prior	no prior
$z_{ik}z_{jk}$	.893$\pm$.002	.898$\pm$.004	.863$\pm$.002	.850$\pm$.002
e^{-d_A}	.958$\pm$.001	.945$\pm$.002	.925$\pm$.001	.908$\pm$.002
$z_{ik}z_{jk}\,e^{-d_A}$	.953$\pm$.002	.943$\pm$.002	.920$\pm$.001	.909$\pm$.002
$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.952$\pm$.001	.940$\pm$.002	.915$\pm$.001	.905$\pm$.002

Table 11: Mechanism ablation on synthetic graphs ($N = 5000$, $K = 4$, mean degree 16), one graph per generating mechanism. Mean±sd over five repetitions, each the best of five random restarts. Aligned: after the best orthogonal map onto the planted coordinates. Top-1: share of held-out edges whose fitted dominant channel matches the generator's.

Generator	Fitted mechanism	ROC ↑	ℓ_1 ↓	ℓ_1 aligned ↓	BNMI ↑	Top-1 ↑
product	$z_{ik}z_{jk}$	.722±.001	.602±.004	.602±.004	.480±.006	.738±.003
	e^{-d_A}	.692±.002	.846±.077	.702±.014	.258±.041	–
	$z_{ik}z_{jk}\,e^{-d_A}$	.692±.002	.743±.068	.690±.007	.297±.043	.633±.073
	$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.677±.004	.878±.067	.753±.027	.211±.034	.473±.070
distance	$z_{ik}z_{jk}$	.813±.001	.706±.038	.613±.013	.458±.016	–
	e^{-d_A}	.845±.000	.708±.150	.475±.002	.511±.049	–
	$z_{ik}z_{jk}\,e^{-d_A}$	.847±.000	.822±.063	.480±.002	.477±.018	–
	$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.838±.002	.611±.082	.501±.004	.530±.030	–
product × full	$z_{ik}z_{jk}$	.857±.000	.396±.001	.396±.001	.781±.001	.922±.001
	e^{-d_A}	.886±.000	.666±.181	.324±.000	.662±.088	–
	$z_{ik}z_{jk}\,e^{-d_A}$	.890±.000	.334±.022	.312±.001	.810±.010	.920±.015
	$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.881±.001	.732±.044	.340±.002	.645±.015	.592±.040
product × complement	$z_{ik}z_{jk}$	.814±.001	.520±.022	.508±.011	.606±.019	.801±.017
	e^{-d_A}	.852±.000	.784±.103	.467±.001	.521±.032	–
	$z_{ik}z_{jk}\,e^{-d_A}$	.855±.000	.884±.076	.475±.001	.488±.012	.455±.069
	$z_{ik}z_{jk}\,e^{-D_{ij,k}}$ (ACE)	.854±.001	.474±.031	.447±.009	.633±.020	.793±.029

K Explanation Analysis on Real Graphs

All statistics use the best $K = 9$ link-prediction fit per graph and its held-out edges, unless stated otherwise. The effective number of channels of an edge is $\exp\big(-\sum_k \omega_k(i,j)\log\omega_k(i,j)\big)$, which is 1 when one channel carries the whole allocation and K when all carry the same.

Table 12: Explanation statistics on held-out edges ($K = 9$, best-objective fit per graph). Null pairs keep one endpoint of a held-out edge and replace the other by one of 10 random nodes, of similar training degree or drawn uniformly. Largest allocations, reassigned allocation and share of the bound are medians; the effective number of channels is a mean. Chance: probability that two held-out edges share the dominant channel.

	Cora	GrQc	HepTh	AstroPh	DBLP
Concentration					
Effective number of channels	2.16	1.78	2.03	2.05	1.73
Largest allocation	0.85	0.91	0.87	0.82	0.92
Same dominant channel by chance	0.11	0.13	0.12	0.12	0.12
Degree-matched null					
Largest allocation of null pairs	0.64	0.64	0.64	0.66	0.65
Real edge sharper than null pairs	0.71	0.87	0.78	0.72	0.87
Same dominant channel as real edge	0.52	0.53	0.50	0.47	0.54
Differs from kept endpoint's usual channel	0.41	0.31	0.35	0.34	0.30
Random-partner null					
Largest allocation of null pairs	0.60	0.62	0.61	0.64	0.62
Real edge sharper than null pairs	0.74	0.87	0.79	0.75	0.88
Same dominant channel as real edge	0.53	0.65	0.58	0.54	0.62
Differs from kept endpoint's usual channel	0.33	0.18	0.26	0.29	0.19
Relation to the Hadamard closure $\mathbf{q}_{ij}$					
Reassigned, $\frac{1}{2}\|\boldsymbol{\omega}_{ij} - \mathbf{q}_{ij}\|_1$	0.055	0.013	0.030	0.029	0.019
Dominant channel changed	0.11	0.05	0.09	0.07	0.05

Table 13: Channels of one held-out edge of Cora at $K = 9$ (nodes 206 and 676, $p = 0.00237$), ordered by allocation. Shared membership $z_{ik}z_{jk}$, complement distance $D_{ij,k}$, the closure $\mathbf{q}_{ij}$ of the shared memberships, the allocation $\boldsymbol{\omega}_{ij}$ and the probability contribution $\pi_{ij,k}$. Among the held-out edges the model predicts and whose leading channel differs from the closure's, this is the one with the median reallocation $\|\boldsymbol{\omega}_{ij} - \mathbf{q}_{ij}\|_1$ among those whose leading channel is ahead of the runner-up by at least $e^{0.4}$.

Channel k	$z_{ik}z_{jk}$	$D_{ij,k}$	$q_{ij,k}$	$\omega_k(i,j)$	$\pi_{ij,k}$
8	0.0186	2.58	0.260	0.604	0.00143
6	0.0391	3.93	0.546	0.329	0.00078
3	0.00667	4.49	0.093	0.032	7.56e-05
4	0.00266	4.51	0.037	0.013	2.98e-05
0	0.0027	4.53	0.038	0.012	2.95e-05
5	0.00111	4.31	0.016	0.006	1.52e-05
1	0.000339	4.37	0.005	0.002	4.34e-06
2	0.000389	4.51	0.005	0.002	4.33e-06
7	1.97e-05	4.53	0.000	0.000	2.15e-07

A worked edge. equation 28 splits the log-ratio of two allocations into a shared-membership term and a complement term, and Table 13 shows both on one held-out edge of Cora. Channel 6 has the larger shared membership, by a factor of 2.1, hence the closure $\mathbf{q}_{ij}$ of the shared memberships gives it 0.55 of the edge against 0.26 for channel 8. The endpoints agree much more closely outside channel 8 than outside channel 6 ($D_{ij,8} = 2.58$ against 3.93), and this reverses the order, since $\log(\omega_8/\omega_6) = -0.74 - (-1.35) = 0.61$ combines -0.74 from the shared memberships with $+1.35$ from the complement distance, such that channel 8 takes 0.60 of the allocation and channel 6 0.33.

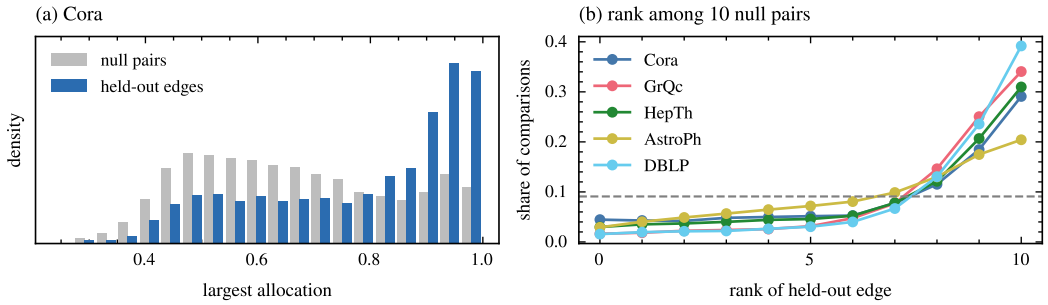


Figure 5: Concentration against degree-matched null pairs. (a) Largest allocation of held-out edges and of null pairs on Cora. (b) Rank of each held-out edge among its ten null pairs (dotted: chance).

The edge is explained by the archetype whose shared membership survives the comparison of everything else, not by shared membership alone. Both channels contribute to the predicted probability, $\pi_{ij,8} = 1.4 \times 10^{-3}$ and $\pi_{ij,6} = 0.8 \times 10^{-3}$ of $p_{ij} = 2.4 \times 10^{-3}$.

Concentration against a null. Null pairs keep one endpoint of a held-out edge and replace the other by one of ten nodes of similar training degree (Table 12, Figure 5). Real edges are sharper. Their median largest allocation is 0.82–0.92 against 0.64–0.66, and a held-out edge is sharper than all ten of its null pairs in 20–39% of comparisons, against 9% by chance. The dominant channel of a null pair matches the real edge's in 47–54% of cases (chance 12–13%), as expected when the kept endpoint belongs mainly to one archetype, but the most common dominant channel among the null pairs differs from the real edge's in 30–41% of cases, so the explanation depends on both endpoints. Uniformly drawn partners give the same conclusions.

Table 14: Links lost when one channel is removed on Cora ($K = 9$, full-graph fit without labels; as many predicted links as observed edges). Majority topic among the endpoints of the lost links, its share among them and among the endpoints of the surviving links, and the share of lost links within one topic. Rows as in Figure 2c.

Channel	Majority topic	Links lost	Share, lost	Share, surviving	Lost within one topic
6	Theory	553	63%	8%	88%
7	Reinforcement learning	676	74%	2%	83%
3	Genetic algorithms	731	83%	8%	86%
4	Neural networks	673	92%	16%	90%
5	Neural networks	551	53%	23%	79%
1	Probabilistic methods	515	42%	12%	94%
8	Probabilistic methods	626	39%	12%	58%
0	Case-based	435	41%	6%	95%
2	Rule learning	542	30%	2%	65%

Lost links on Cora. Table 14 gives the numbers behind Figure 2c. The labelled graph is fitted on all its edges without labels, all pairs are scored, and τ is set so that the predicted links equal the observed edges in number.

Table 15: Links lost when one channel is removed, per labelled graph (full-graph fits without labels; as many predicted links as observed edges). Share of the majority class among the endpoints of each channel's lost links and among those of its surviving links: median over channels (range). Enriched: channels whose majority class has a larger share among lost than among surviving links.

Graph	K	Predicted links	Observed	Within one class	Classes covered	Share, lost	Share, surviving	Enriched
Cora	9	5,069	61%	82%	7/7	53% (30%–92%)	8% (2%–23%)	9/9
Cora	17	5,069	70%	81%	7/7	55% (30%–94%)	13% (3%–27%)	17/17
Citeseer	17	3,668	70%	74%	5/6	64% (36%–81%)	18% (5%–40%)	16/17
LastFM	17	27,806	62%	92%	10/18	70% (28%–98%)	10% (0%–19%)	17/17

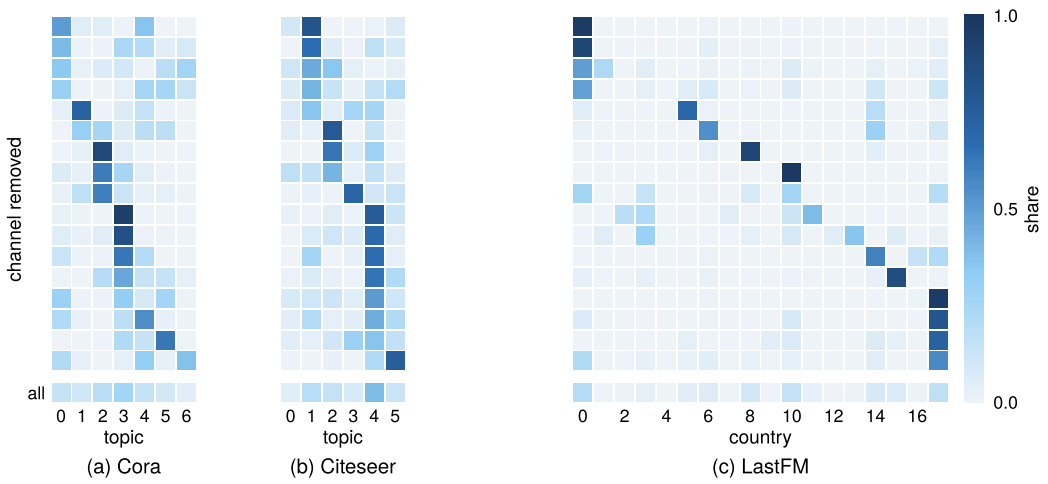


Figure 6: Class shares among the endpoints of predicted links lost when one channel is removed, at $K = 17$ on the three labelled graphs (full-graph fits without labels; one row per channel, ordered by majority class; bottom row: all predicted links).

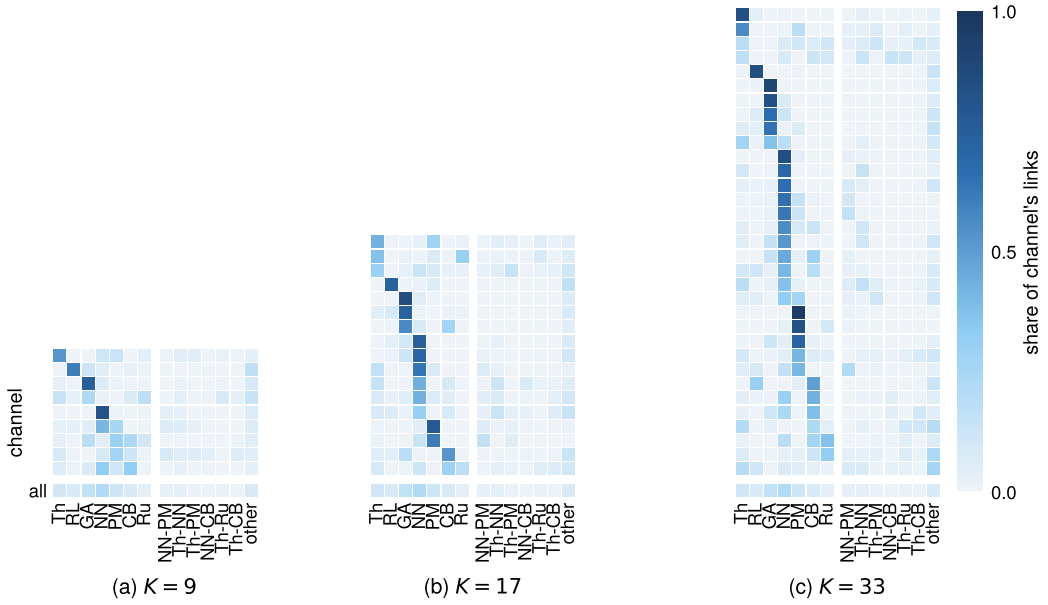


Figure 7: Topic-pair composition of each channel on Cora at $K = 9$, 17 and 33. Rows: channels, ordered by their dominant pair; within-topic pairs left of the gap; bottom row: all edges.

Lost links on all labelled graphs. Figure 6 and Table 15 repeat the analysis on Cora, Citeseer and LastFM at $K = 17$. On every graph the links lost when one channel is removed are dominated by one class, whose share among their endpoints is 30–94% on Cora, 36–81% on Citeseer and 28–98% on LastFM, against 3–27%, 5–40% and 0–19% among the links that remain, and this class is over-represented among the lost links for all but one of the 51 channels. The channels cover all seven Cora topics, five of six Citeseer topics and ten of the LastFM countries (Table 15).

Topic pairs by channel. To see what each channel connects, we weight the topic pair of every observed Cora edge by its allocation, at $K = 9$, 17 and 33 (Figure 7, Table 16). Although 80% of Cora's edges join nodes of the same topic, so that every channel is mostly within-topic, each channel concentrates on one particular topic, at a median 3.5–3.7 times that pair's share of all edges. As K

Table 16: What each channel connects on Cora (full-graph fits without labels). Per K: channels whose dominant topic pair lies within one topic, topics covered by them, and medians over channels of the dominant pair's share of the channel's links, its lift over the pair's share of all edges, and the channel's within-topic share (all edges: 0.80). Parent: the most similar channel at the next smaller K; similarity is the cosine of their allocations over the same edges, and parent topic counts the channels whose dominant pair is their parent's.

K	Within-topic	Topics	Dominant share	Lift	Within share	Parent sim.	Parent topic
9	9/9	6	0.41	3.7	0.83	–	–
17	17/17	6	0.57	3.5	0.82	0.36	14/17
33	33/33	7	0.52	3.5	0.84	0.37	21/33

grows the channels are rebuilt rather than split, with a median similarity of 0.36–0.37 to the closest channel at the next smaller K, although most keep that channel's topic (14 of 17 and 21 of 33).

L Relational Structure Across Membership Ranks

For node i, let $k_i^{(r)}$ denote the archetype occupying the rth largest component of $\mathbf{z}_i$. At each rank we quantify within- versus between-archetype connectivity by

$$\eta_r = P(Y_{ij} = 1 \mid k_i^{(r)} = k_j^{(r)}) / P(Y_{ij} = 1 \mid k_i^{(r)} \neq k_j^{(r)}).$$

To separate lower-rank structure from associations induced by dominant membership, we stratify node pairs by the unordered dominant-archetype pair $\{k_i^{(1)}, k_j^{(1)}\}$ and compute a Mantel–Haenszel adjusted risk ratio. As a conditional null, for $r > 1$ we permute $k_i^{(r)}$ within groups sharing the same dominant archetype. This preserves the association between lower-rank and dominant memberships while removing residual association with connectivity. Figure 4 reports the raw enrichment, the dominant-pair-adjusted enrichment, and the 95% permutation-null interval across ranks.

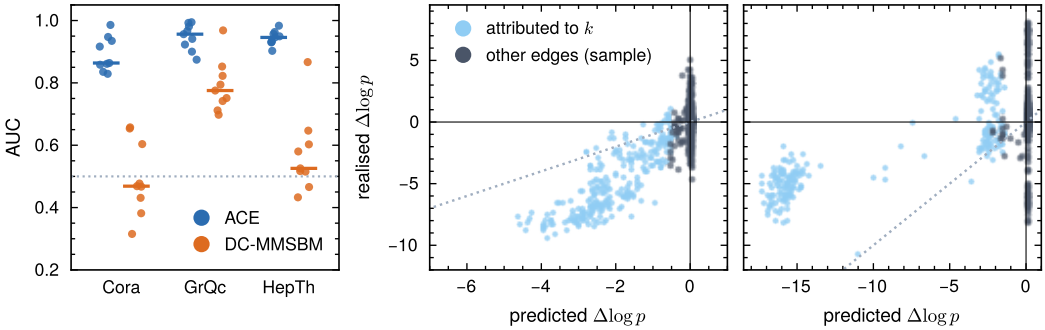
M Counterfactuals Against Refitting

For each of the nine channels of the best $K = 9$ fit on Cora, GrQc and HepTh, the training edges whose dominant channel is k are removed and the model is refitted three times from scratch with the same protocol, keeping the fit with lowest training objective. After removal these pairs are no longer training edges, and therefore the refitted model may sample them as non-edges and sees them as zeros rather than missing under the convention of Appendix A. As a control, the same number of random training edges is removed and the model refitted in the same way. Held-out edges are never removed, and all changes are measured on them against the original fit. Hit edges are the 10% of held-out edges whose log-probability the refit lowers most. The closed form assigns each held-out edge the change in log-probability that Theorem 3.3 predicts when the channel is removed from the original fit. The DC-MMSBM is tested in the same manner with its own attribution, the largest role marginal of the edge posterior, and its closed form sets the removed role's row and column of the block matrix to zero.

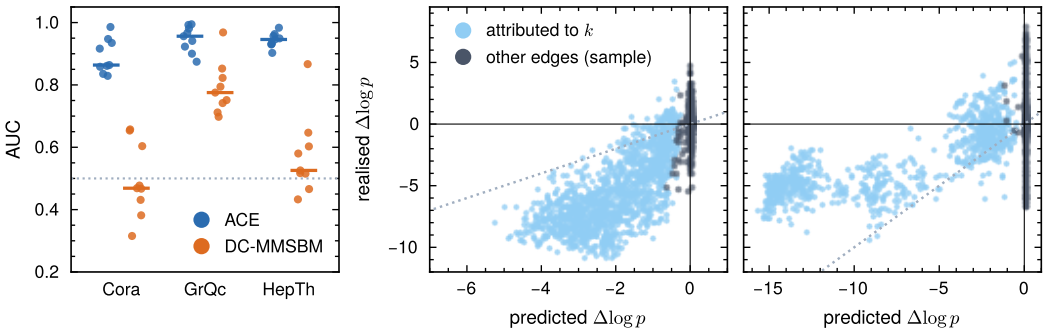
Results. Table 17 reports the medians over channels or roles. The results show that ACE's refits rarely contradict the closed form. Only 2–3% of attributed edges gain probability, and at most 0.5% of the other held-out edges move by more than 5 in log-probability, against 9.5–17% and 9–16% for the DC-MMSBM. The closed form ranks the size of the change only moderately (per-edge Spearman correlation 0.25–0.28), hence it identifies which edges are affected more reliably than by how much. The DC-MMSBM's closed form predicts a drop of 7.2–10.8 on attributed edges against a realised drop of 2.9–4.7, a collapse the refit does not deliver. Figure 8 shows the Cora and HepTh versions of Figure 3.

Table 17: Counterfactuals against refitting ($K = 9$). For each channel (ACE) or role (DC-MMSBM), the training edges attributed to it are removed and the model refitted; changes in $\log p$ are measured on held-out edges, which are never removed. Hit edges: the 10% of held-out edges the refit lowers most. Median over the nine components of each graph; parentheses: range over components.

	Cora	GrQc	HepTh
ACE			
Held-out edges attributed to k, after refit	−4.51	−6.92	−5.53
Other held-out edges, after refit	−.064	−.020	−.020
Attributed edges, random removal	−.532	−.452	−.530
Other edges, random removal	−.595	−.462	−.518
Attributed edges, closed form	−2.03	−2.11	−2.06
AUC, closed form identifies hit edges	.864 (.829–.986)	.956 (.874–.995)	.946 (.903–.983)
AUC, ranked by change under random removal	.580	.546	.554
Hit edges attributed to k	.736	.841	.812
Attributed edges whose probability rises	.019	.019	.032
Other edges moving by more than 5	.001	.002	.005
Spearman, closed form vs refit	.279	.253	.265
DC-MMSBM			
Held-out edges attributed to k, after refit	−2.88	−4.67	−3.08
Other held-out edges, after refit	.027	−.083	−.119
Attributed edges, closed form	−9.38	−10.81	−7.21
AUC, closed form identifies hit edges	.469 (.315–.657)	.775 (.698–.968)	.526 (.433–.867)
Hit edges attributed to k	.319	.529	.287
Attributed edges whose probability rises	.167	.095	.142
Other edges moving by more than 5	.160	.088	.099



(a) closed form identifies the affected edges (b) ACE, Cora, channel 7 (c) DC-MMSBM, Cora, role 8



(a) closed form identifies the affected edges (b) ACE, HepTh, channel 6 (c) DC-MMSBM, HepTh, role 6

Figure 8: As Figure 3, for Cora (top) and HepTh (bottom).

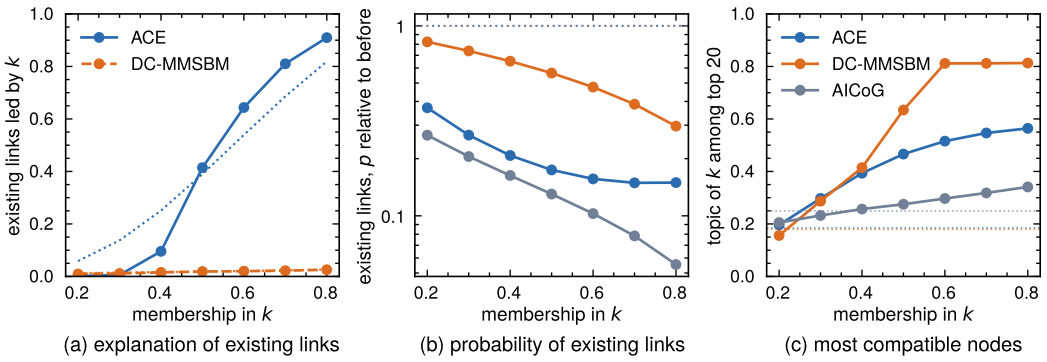


Figure 9: The three models at matched membership, for topic-pure archetypes the node starts below 0.15. (a) Share of the node's existing links whose dominant channel (ACE) or role (DC-MMSBM) is the targeted archetype, with the mean allocation to it dotted. (b) Probability of the existing links relative to before the intervention. (c) Targeted topic among the 20 most compatible candidates of degree at most five, ranked without activities, with base rates dotted.

N MEMBERSHIP INTERVENTIONS

We use the full-graph fits of labelled Cora at $K = 9$, fitted without labels. The nodes are split at random into two halves. Each archetype is assigned the majority topic of the first-half nodes it dominates; the interventions, their candidate links and each archetype's purity (the share of the second-half nodes it dominates that carry its topic) use the second half. We select 300 second-half nodes of degree at least two and shift each towards every archetype whose topic differs from its own, along equation 6 with $t \in \{0, 0.5, \ldots, 4\}$. This gives 2,201 interventions, 1,220 of them towards topic-pure archetypes, whose purity exceeds one half.

Table 18: Membership interventions on labelled Cora ($K = 9$, full-graph fits without labels). Top: Theorem 3.2 and Equation 8 for ACE. Middle: topics of the node's 20 most probable new links before ($t = 0$) and after ($t = 4$) the intervention; topic-pure archetypes are those whose dominant nodes are mostly one topic. Bottom: the three models at matched membership $z_{ik} = 0.8$ after the intervention, for topic-pure archetypes the node starts below 0.15, with candidates of degree at most 5 ranked without activities.

Theorem 3.2 and Equation 8 (ACE)				
Interventions (node, archetype)		2,201 (1,220 topic-pure)		
Link-steps checked against Equation 7, violations		61,116, 0		
Edges with target inside the bracket ($w^\star = 0.5$)		3,406 of 3,406		
Topics of the most probable new links (ACE)				
Archetypes	Target topic, $t = 0$	Target topic, $t = 4$	Base rate	Own topic, $t = 0 \to 4$
Topic-pure ($n = 1{,}220$)	8.9%	49.2%	18.4%	60.7% $\to$ 28.5%
Other ($n = 981$)	7.3%	21.0%	17.8%	60.3% $\to$ 34.7%
All ($n = 2{,}201$)				60.5% $\to$ 31.3%
Matched membership $z_{ik} = 0.8$, 1,029 candidates				
Model	Existing links now led by k	Allocation to k	Target topic (base rate)	Median t
ACE	91.0%	81.9%	56.4% (18.5%)	4.9
AICoG	–	–	34.1% (24.9%)	6.9
DC-MMSBM	2.6%	2.7%	81.3% (18.0%)	10.5

Theorem 3.2 and predictions. On every existing link of every intervened node, the allocation log-odds of the targeted channel rise between consecutive perturbations at a rate within the bounds of equation 7, with no violation in 61,116 link-steps, and the bracket of equation 8 contains the perturbation needed for an allocation of 0.5 on all 3,406 edges checked (Table 18). The redirection of predicted links is weaker for archetypes that are not topic-pure, whose topic's share rises from 7.3% to 21.0% (base rate 17.8%), consistent with their weaker alignment with a single topic.

Comparison with AICoG and the DC-MMSBM. Any compositional model can be intervened on. To compare the models at the same change in membership, we raise a node's membership in a topic-pure archetype it starts below 0.15 to $z_{ik} = w$, and rank candidates of degree at most five by the

membership part of each model's score, without activities, so that popularity plays no role (Figure 9). All three models redirect the most compatible nodes towards the archetype's topic. At $z_{ik} = 0.8$, the DC-MMSBM does so most (81% of the top 20, against 56% for ACE and 34% for AICOG), but it needs a far larger perturbation to reach that membership (median $t = 10.5$, against 4.9 for ACE and 6.9 for AICOG). Only ACE's explanation of the node's existing links follows the intervention. The targeted archetype becomes the dominant channel on 91% of them for ACE, against 3% for the DC-MMSBM, whose role posterior barely moves (mean allocation 2.7%, against 82% for ACE); AICOG has no per-edge allocation. Any compositional model can thus redirect predictions, but only ACE states in advance how the intervention changes each edge's explanation, and by how much.