Mathematical Model: Graphs as Vectors and Semantic Projections

Rahul Krishna Gaddam

September 23, 2025

1 Objects and Base Space

Let the universe of possible node types be

$$V = \{v_1, \dots, v_M\}.$$

We treat each node type v_i as a canonical coordinate (basis vector) in $\mathbb{R}^M$. A graph G concerning a particular customer is represented by:

- a node-activity vector $x_G \in \mathbb{R}^M$, where $x_G[i]$ represents the presence/strength/count/importance of node-type v_i in G.
- a relation matrix $A_G \in \mathbb{R}^{M\times M}$ whose (i, j) entry $A_G[i, j]$ is the weight of the relation from v_i to v_j. For undirected graphs, A_G is symmetric.

Thus, the raw representation of a graph is the pair (x_G, A_G), which is a point in

$$X = \mathbb{R}^M \times \mathbb{R}^{M\times M} \cong \mathbb{R}^{M+M^2}.$$

2 Linearized Vector Embedding

Flatten A_G column-major to $\mathrm{vec}(A_G) \in \mathbb{R}^{M^2}$. Define the linear embedding:

$$\Phi_{\mathrm{lin}}(G) = \begin{bmatrix} x_G \\ \mathrm{vec}(A_G) \end{bmatrix} \in \mathbb{R}^D, \quad D = M + M^2.$$

This is a full, lossless linear encoding of the graph (for fixed M).

3 Learnable Projection to Semantic Space

We want embeddings in $\mathbb{R}^d$ with $d \ll D$. Introduce a learnable map $W \in \mathbb{R}^{d\times D}$ and bias b:

$$e_G = f(W\Phi_{\mathrm{lin}}(G) + b) \in \mathbb{R}^d,$$

where $f(\cdot)$ is an optional nonlinearity (e.g., tanh).

4 Normal Plane / Projection Idea

Suppose there exists a meaningful subspace $S \subseteq \mathbb{R}^d$ (the *semantic plane*) capturing invariant customer attributes. The orthogonal projector onto S is $P_S \in \mathbb{R}^{d\times d}$ with

$$P_S^2 = P_S, \quad P_S^\top = P_S.$$

For any embedding e:

$$\mathrm{proj}_S(e) = P_S e.$$

If two versions G_1, G_2 differ only by orthogonal changes:

$$P_S e_{G_1} = P_S e_{G_2}.$$

Equivalently, $\Delta = e_{G_1} - e_{G_2} \in S^\perp$.

5 Retrieval Objective and Minimal Flow

5.1 Cosine Similarity and Residual Penalty

We define similarity:

$$\mathrm{sim}(e_a, e_b) = \frac{e_a^\top e_b}{\|e_a\|\|e_b\|}.$$

Projection-based retrieval score:

$$\mathrm{score}(G_q, G_c) = \alpha \cdot \frac{(P_S e_{G_q})^\top (P_S e_{G_c})}{\|P_S e_{G_q}\|\|P_S e_{G_c}\|} - \beta \cdot \|(I - P_S)(e_{G_q} - e_{G_c})\|,$$

with $\alpha, \beta \geq 0$.

5.2 Minimal Flow (Optimal Transport View)

Interpret each graph version as a probability distribution p_G from x_G. Define Earth Mover's Distance (EMD):

$$\mathrm{EMD}(p_{G_1}, p_{G_2}) = \min_{T\geq 0} \sum_{i,j} T_{ij} c_{ij},$$

subject to row/column marginals matching p_{G_1}, p_{G_2}.

Alternatively, minimal perturbation in embedding space:

$$\delta^* = \arg\min_\delta \|\delta\| \quad \text{s.t.} \quad e_{G_1} + \delta = e_{G_2}.$$

Hence minimal flow norm is $\|e_{G_2} - e_{G_1}\|$.

6 Theorems and Lemmas

Lemma (Projection Invariance). If $e_{G_2} = e_{G_1} + \delta$ with $\delta \in S^\perp$, then retrieval using only $P_S e$ treats G_1, G_2 as identical.

Spectral Stability. If e_G is derived from Laplacian eigenmaps, perturbations to edges/nodes yield bounded changes in eigenvectors (Davis–Kahan theorem).

7 Training Objectives

7.1 Supervised Contrastive Objective

For positives G^+ (same customer) and negatives G^-:

$$L_{\text{NCE}} = -\sum_q \log \frac{\exp(\text{sim}(P_S e_q, P_S e_{q^+})/\tau)}{\sum_{c\in\text{batch}} \exp(\text{sim}(P_S e_q, P_S e_c)/\tau)}.$$

7.2 Learnable Projection

Parameterize

$$e_G = U z_G, \quad U \in \mathbb{R}^{d\times k},\ z_G \in \mathbb{R}^k,$$

where $P_S = UU^\top$. Optimize jointly with contrastive loss.

7.3 Residual Consistency Regularizer

$$L = L_{\text{NCE}} + \lambda \sum_{(G_i,G_j)\in\text{same-id}} \|(I - P_S)e_{G_i} - (I - P_S)e_{G_j}\|^2.$$

8 Concrete Computable Formulas

- Build x_G and $\text{vec}(A_G)$, concatenate into $\Phi_{\text{lin}}(G)$.
- Learn $W \in \mathbb{R}^{d\times D}$ and compute $e_G = W\Phi_{\text{lin}}(G)$.
- Estimate S via PCA (option 1) or learnable U (option 2).
- Retrieval score:

$$\text{score}(G_q, G_c) = \frac{(P_S e_q)^\top (P_S e_c)}{\|P_S e_q\|\|P_S e_c\|} - \gamma\|(I - P_S)(e_q - e_c)\|.$$

9 Example Sketch

Let $M = 3$ (Name, Address, Phone). Suppose:

$$x_{G_1} = [1, 1, 1]^\top, \quad x_{G_2} = [1, 1.1, 0.9]^\top.$$

After projection, $P_S e_{G_1} \approx P_S e_{G_2}$, showing invariance to small variations (e.g. phone).

10 Implementation Checklist

1. Choose node types V.
2. Encode via $\Phi_{\text{lin}}(G)$.
3. Learn embedding e_G.
4. Extract semantic subspace S.
5. Use projection-based retrieval.
6. Optionally compute EMD for minimal flow.

11 Proof Sketch

If true semantic identity corresponds to $s^\top e$, where $s \in S$, then

$$s^\top e = s^\top P_S e,$$

and any orthogonal noise $\delta \in S^\perp$ vanishes: $s^\top \delta = 0$.

Thus projection ensures robust similarity.

12 Hackathon Wrap-Up

- Equations + visualization of e_G, $P_S e_G$, residuals.
- Demo: 3 versions of same customer $\rightarrow$ naive system = 3 records, projection system = 1 cluster.
- Training: show contrastive loss enforcing projection invariance.