Emergent Knowledge Graphs from High-Order Semantic Spaces

Patrick Audley
Blackcat Informatics Inc.

October 2025

Abstract

This paper describes a reproducible, from-first-principles method for discovering emergent knowledge-graph relations in high-order semantic vector spaces. By combining manifold learning, spectral analysis, and residual factorization, we demonstrate how to extract non-linear conceptual structures from embedded text corpora. The workflow can be reimplemented from the mathematical and procedural definitions provided here, without reliance on any specific codebase. Outputs include interpretable graphs, cosine-based similarity measures, and labeled semantic relations suitable for automated ontology construction.

1 Introduction

Semantic embeddings transform text into points in a high-dimensional vector space in which proximity encodes meaning [Mikolov et al., 2013, Devlin et al., 2019]. Dimensional-reduction and graph-learning techniques, such as Uniform Manifold Approximation and Projection (UMAP) [McInnes et al., 2018], can reveal the geometric structure of this space by modeling its local neighborhoods. Principal Component Analysis (PCA) [Hotelling, 1933] isolates the dominant linear modes of variance, while the residual components retain the manifold's non-linear relationships.

By constructing spectral embeddings of both the original and residual spaces, we can visualize and quantify how semantic curvature generates conceptual bridges between distinct clusters of meaning. These bridges correspond to knowledge-graph links inferred directly from the geometry of language representations.

We are posting this paper, which represents the start of our explorations in the field, in order to spark conversation. This work hints at a much larger field that we are actively exploring.

2 Corpus and Data Preparation

2.1 Text Corpus

We begin with a set of ten thematically related short essays, each representing a distinct but overlapping domain (e.g., renewable energy, ecology, urban planning, Indigenous stewardship). Each document is preprocessed into smaller "semantic chunks" by greedily concatenating sentences until the text length lies within $[250, 800]$ characters.

2.2 Vector Embedding

Each chunk is embedded into a d-dimensional semantic space using a pre-trained language model (e.g., a Gemini or BERT-style embedding function $f : T \to \mathbb{R}^d$). For each text t_i we obtain an embedding vector

$$x_i = \frac{f(t_i)}{\|f(t_i)\|_2}, \tag{1}$$

ensuring that cosine similarity and dot product coincide.

3 Linear and Nonlinear Factorization

3.1 Principal Component Decomposition

Given the embedding matrix $X \in \mathbb{R}^{n \times d}$, PCA identifies orthogonal basis vectors W_k that maximize variance:

$$X_{PCA} = XW_k, \qquad \hat{X} = X_{PCA}W_k^\top. \tag{2}$$

$\hat{X}$ represents the best k-dimensional linear approximation of X. These directions correspond to global topical axes in the corpus.

3.2 Residual Space Construction

The residual embeddings are computed as the component of each vector not explained by the linear model:

$$R = X - \hat{X}, \qquad R_i \leftarrow \frac{R_i}{\|R_i\|_2}. \tag{3}$$

This isolates the non-linear structure of the embedding manifold.

This approach is mathematically adjacent to the "All-but-the-Top" (ABTT) method proposed by Mu et al. [2018], which removes the top principal components from word embeddings to eliminate common discourse patterns and improve semantic similarity measures. While ABTT focuses on postprocessing static word vectors to enhance their quality, our residual space construction (which we term the *EigenDARK* procedure) explicitly separates linear and non-linear semantic structures to discover emergent knowledge-graph relations that are not captured by first-order linear projections.

4 Manifold Learning and Graph Construction

4.1 Spectral UMAP Embedding

UMAP constructs a weighted neighborhood graph $G = (V, E)$ from pairwise distances $d(x_i, x_j)$ using the metric $\text{cosine}(x_i, x_j) = 1 - x_i \cdot x_j$. For each point i, a local connectivity scale σ_i is found such that

$$\sum_j \exp\left(-\frac{\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2(k), \tag{4}$$

where k is the desired number of neighbors. The symmetric fuzzy graph weights are then

$$\begin{aligned} w_{ij} = \exp\left(-\frac{\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) + \exp\left(-\frac{\max(0, d(x_j, x_i) - \rho_j)}{\sigma_j}\right) \\ - \exp\left(-\frac{\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) \exp\left(-\frac{\max(0, d(x_j, x_i) - \rho_j)}{\sigma_j}\right). \end{aligned} \tag{5}$$

The 2D coordinates y_i are initialized via the first two eigenvectors of the graph Laplacian $L = D - A$, where $A = [w_{ij}]$, following Laplacian eigenmaps [Belkin and Niyogi, 2003].

4.2 Dual-UMAP Analysis

Two embeddings are computed:

1. $Y_{orig} = \text{UMAP}(X)$, the projection of the full semantic manifold.
2. $Y_{resid} = \text{UMAP}(R)$, the projection of the residual (non-linear) space.

Edges and clusters in Y_{resid} represent relationships that are not linearly recoverable.

5 Quantifying Nonlinear Relationships

For each high-confidence UMAP edge (i, j) with weight w_{ij}, we compute three cosine similarities:

$$c_{orig} = x_i \cdot x_j, \tag{6}$$

$$c_{pca} = \hat{x}_i \cdot \hat{x}_j, \tag{7}$$

$$c_{res} = r_i \cdot r_j. \tag{8}$$

We define the *nonlinear surprise* as

$$\Delta_{nl} = c_{res} - c_{pca}. \tag{9}$$

Large positive Δ_{nl} values indicate pairs of points whose affinity increases when linear structure is removed, implying a nonlinear semantic connection.

6 Semantic Edge Labeling

To interpret the meaning of strong edges, we compute direction vectors $v_{ij} = (x_j - x_i)/\|x_j - x_i\|$ and compare them with probe vectors p_t derived from key terms. The alignment score

$$s_t = p_t \cdot v_{ij} \tag{10}$$

measures how closely the semantic direction matches each probe concept. The top-k aligned probes yield natural-language descriptors for the relation (e.g., "policy implementation," "biodiversity resilience").

7 Visualization and Interpretation

Edge	c_{orig}	c_{pca}	Δ_{nl}
(Indigenous, Energy)	0.72	0.45	+0.27
(Agriculture, Water)	0.68	0.61	+0.07
(Mining, Ecology)	0.53	0.58	-0.05

Table 1: Example edge metrics showing nonlinear surprise across semantic domains.

8 Reproduction Procedure

A researcher can reproduce identical results by following these steps:

Figure 1: UMAP projection of the original embedding space. Edges are colored by nonlinear surprise Δ_{nl}. Red edges indicate nonlinear conceptual bridges.

1. Prepare ten domain texts and segment them into 250–800 character chunks.
2. Compute normalized embeddings x_i using any high-quality transformer embedding model.
3. Perform PCA on X, compute reconstructions $\hat{X}$ and residuals R.
4. Build neighborhood graphs and run spectral UMAP on both X and R with parameters ($n_{\text{neighbors}} = 12, \text{min\_dist} = 0.08$).
5. Compute edge cosines $c_{orig}, c_{pca}, c_{res}$ and nonlinear surprise Δ_{nl}.
6. Project the UMAP graph to 2D, color edges by Δ_{nl}, and visualize clusters.
7. Derive semantic labels by embedding a vocabulary of candidate terms and comparing them to edge direction vectors.
8. Tabulate all metrics and labels into a CSV for further analysis.

9 References

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

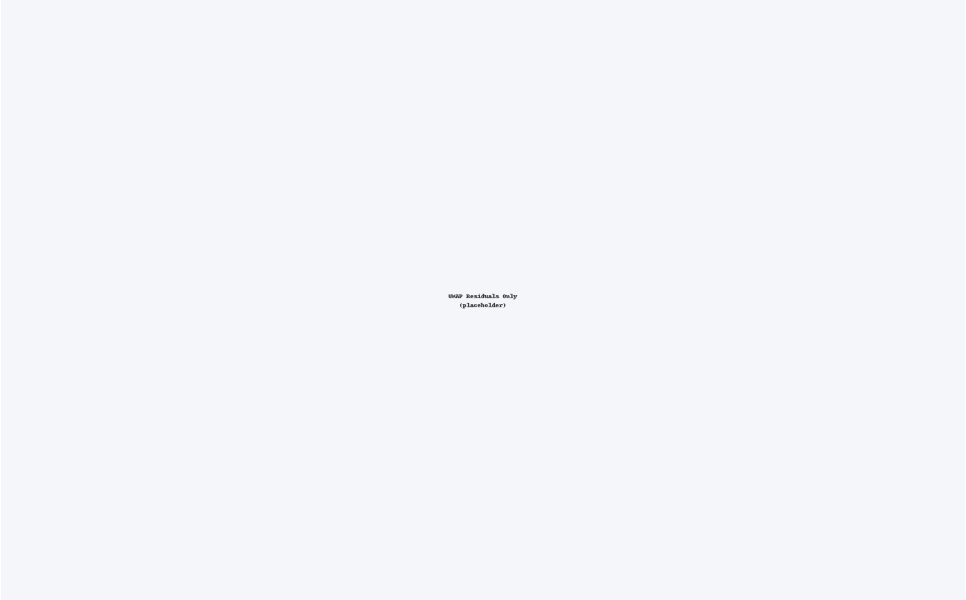

Figure 2: UMAP projection of the residual space, isolating nonlinear semantic relationships.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mu, J., Bhat, S., and Viswanath, P. (2018). All-but-the-Top: Simple and Effective Post-processing for Word Representations. arXiv:1702.01417.