Explaining Graph Embeddings. Zohair Shafi, Ayan Chatterjee, Tina Eliassi-Rad. Northeastern University.

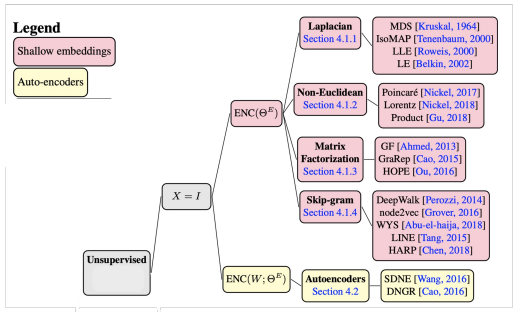
Graph Embedding Problem

Inputs

- Graph G=(*V*, *E*, *X*)
  - Vertex set *V*
  - Edge set *E*
  - Node-attribute matrix *X*
- Size of embedding *d* (*d* << |*V*|)

Output

- Embedding matrix *D* of size |*V*| x *d*



Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C. and Murphy, K., 2020. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. arXiv preprint arXiv:2005.03675.

Research Question #1: *What does each dimension mean? Can we translate them into something a human would understand?*

Proposed Method:[1]

- Step 1: Run a graph embedding method to get the embedding matrix *D.*
- Step 2: Compute a set of *f* topological "sense" features on the nodes to get a feature matrix *F.*
- Step 3: Employ non-negative matrix factorization to find an "explain" matrix *E*: *DE = F*

Outcome: Each row of the explain matrix describes every dimension as a mixed membership of the sense features.

Research Question #2: *Can we explain the embedding of a particular node to a human?*

Proposed Method:

- Let $\vec{y}_i$ and $\vec{f}_i$ be the embedding and sense feature vectors for node i, respectively.
- Compute node i's explain matrix:

$$E_i = \frac{\vec{y}_i^{\mathrm{T}} \vec{f}_i}{\|\vec{y}_i\| \|\vec{f}_i\|}$$

Outcome: Each entry (j, k) in E_i describes how much sense feature k explains the placing of node i in embedding dimension j.

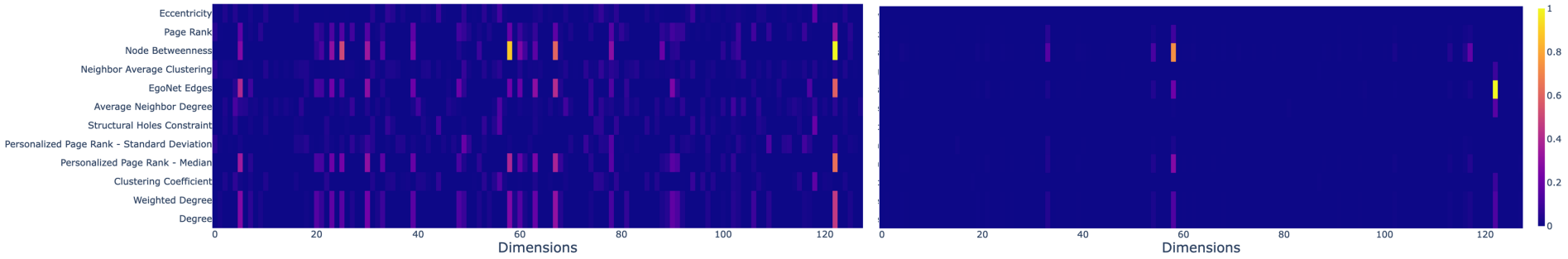
Research Question #3: *How can we modify an existing objective function for graph embedding to improve explainability?*

Proposed Method:

- Augment the objective function of an existing graph embedding method by adding two additional constraints:
  - Require the rows of *E* to be orthogonal to each other. We want low overlap in what each dimension explains.
  - Require the columns of *E* to be sparse. We want each sense feature to contribute to the explanations of as few dimensions as possible.
- To speed up optimization, find the weights that optimize the original objective function and use them as initialize weights for the augmented objective function.

Transpose of the Explain matrix of SDNE[2] on a collaboration network (link prediction AUC = 0.9)

Transpose of the Explain matrix of SDNE+ on a collaboration network (link prediction AUC = 0.87



Our augmented SDNE (a.k.a. SDNE+) improves the explainability of the explain matrix by adding sparsity and orthogonality constraints to SNDE's objective function. It also shows that most of the 128 embedding dimensions were not needed to embed this academic collaboration network.

Take-away Points

- Sense-making explains each dimension in terms of human-understandable features.
- Adding constraints to the explain matrix improves explainability with little impact on performance of downstream tasks.

[1] This method is similar to Henderson et al. KDD 2012.
[2] D. Wang et al. KDD 2016.