Paper Review For JDHMT
by Anurag Banerjee

High quality gene/disease embedding in a multi-relational heterogeneous graph after a joint matrix/tensor decomposition[1]

Kaiyin Zhou et. al.

1 Jargon Glossary

Epigenomic - Describes changes in regulation of gene activities that act independently of changes in gene sequences

Decomposition - Factorisation of a matrix or tensor into simpler terms

Heterogeneous graph - A graph that may contain more than one type of node and edge

2 Problem Description

In this paper, the authors have attempted to achieve the following three objectives:

1. Prepare a format for the representation of the data (called *triples*). In essence, it denotes the dimensions for the initial tensors
2. Whereas the format above is used for storing and manipulating the knowledge graph, the actual information was curated from 7 mainstream datasets. A large-scale heterogeneous *gene-disease* network was thus constructed [2]
3. Embeddings for genes and diseases were generated using a **J**oint **D**ecomposition of **H**eterogeneous **M**atrix and **T**ensor optmisation model. The *principal components* thus obtained form the embeddings

Further, to evaluate the usefulness of the generated embeddings, two approaches were undertaken - *intrinsic*, where, goodness was assessed visually through t-SNE plots and *extrinsic*, where, goodness was assessed by utilising the embeddings in some downstream task.

3 Problem Relevance

In literature, three techniques for generating embeddings from graphs are commonly discussed:

- **Proximity Preserving** - *where local topological information is utilised, e.g., Random Walk, etc.*
- **Message Passing** - *pertains mostly to graph neural networks, e.g., R-GCN, etc.*
- **Relation Learning** - *leverages edge properties, e.g., Singular Value Decomposition, etc.*

The authors have noted that Proximity Preserving paradigms are not very good at capturing global structure information, however, they have not explicitly identified shortcomings for the Message Passing paradigm. They have identified two **triplet formats** where, "*The motive of this research is to perform triplet data fusion in a multi-relational heterogeneous graph*".

4 Proposed Solution

The overall idea of the concept presented in this paper can be summarised in the *mindmap* in Figure 2 and the pictorial representation of the core logic can be seen in Figure 1. The authors have proposed two **triplet formats**:

1. `(gene/disease, uni-relation, heterogeneous entity)`
2. `(gene/disease, multi-relation, gene/disease)`

Here, a **heterogeneous entity** can be anyone of *chemical, mutation, pathway and phenotype*. For the **uni-relation** types, there are 11 alternatives:

1. disease-chemical edge *obtained from CTD*
2. gene-chemical edge *obtained from CTD*
3. disease-pathway edge *obtained from CTD*
4. gene-pathway edge *obtained from CTD*
5. gene-disease edge *obtained from CTD*
6. disease-mutation edge *obtained from DisGeNET*
7. gene-mutation edge *obtained from DisGeNET*
8. disease-phenotype edge *obtained from HPO*
9. gene-phenotype edge *obtained from HPO*
10. gene-gene edge *obtained from BioGRID*
11. disease-disease edge *obtained from BioSNAP*

Similarly, for **multi-relation** there are multiple options (*apparently*), but, for the example described in the paper, they have selected (from **AGAC** database) *mutation type* as the multi- relation, where, it has three variants, *viz.*:

1. loss-of-function mutation
2. gain-of-function mutation
3. complex mutation

Along with **AGAC**, 7 data-sources are used, namely, **BioGRID**, **BioSNAP**, **CTD**, **DisGeNET**, **Disease-Ontology**, **HPO** and **NCBI**. Also, a step called *concept normalisation* is performed where for each node type, different concept IDs are used, such as, **Entrez ID for *gene***, **MESH ID for *disease***, **SNPID for *mutation***, **KEGG ID for *pathway*** and **HPO ID for *phenotype***.

> This paper uses something equivalent of learnt tensor/matrix decomposition!

Further, the authors have noted that, "*...matrix is natural to store uni-relational triple information, while tensor serves well for multi-relational triple*".Thus, the matrix $\mathcal{W} \in \mathbb{R}^{n \times h}$ for *uni-relations* (with rows as gene/disease and coloumn as entity) and the sparse tensor $\mathcal{X} \in \mathbb{R}^{n \times n \times K}$ for *multi-relations* is built from the 7 data-sources for each of the **two** *triplet* formats.
where,
$\mathcal{G}$: Number of unique genes,
$\mathcal{D}$: Number of unique diseases,
K: Number of types for the multi-relation. In the example used, it is 3,
$n = \mathcal{G} + \mathcal{D}$,
$h = \mathcal{G} + \mathcal{D} + \mathcal{M}$,
The values in $\mathcal{W}$ are binary, and those in $\mathcal{X}$ are linkage weights.

The solution proposed in this paper, for generating learned embeddings, is via a joint decomposition of the matrix $\mathcal{W}$ and the sparse tensor $\mathcal{X}$. The computational algorithm (**Algorithm** 1) captures the solution (*the learnt joint-decomposition*):
In **Algorithm** 1, the symbols have the following meaning,
$\mathcal{J}(\theta) = f_{\mathcal{X}}(\mathcal{X}, A, \mathcal{C}) + f_{\mathcal{W}}(\mathcal{W}, A, V)$: The objective function for *convex optimization*
$f_{\mathcal{X}}(\mathcal{X}, A, \mathcal{C}) = \lambda_{\mathcal{X}} \sum_{k=1}^{K} ||\mathcal{X}_k - A\mathcal{C}_k A^T||_F^2 + \lambda_{\mathcal{C}} \sum_{k=1}^{K} ||\mathcal{C}_k||_F^2$: need to understand!
$f_{\mathcal{W}}(\mathcal{W}, A, V) = \lambda_{\mathcal{W}} ||\mathcal{W} - AV||_F^2 + \lambda_A ||A||_F^2 + \lambda_V ||V||_F^2$: need to understand!
$\theta = \{A, V, \mathcal{C}\}$: The parameter-set
$A \in \mathbb{R}^{n \times e}$: need to understand!
$V \in \mathbb{R}^{e \times h}$: need to understand!
$\mathcal{C}_k = \mathcal{C}_{::k} \in \mathbb{R}^{e \times e}$: need to understand!
All values in the matrix and the tensor are strictly non-negative.

> Since my study pertains to understanding the embedding generation, the intrinsic and extrinsic evaluations are not discussed.

5 Positive Points

- The learnt decomposition approach to embedding learning seems promising
- The computation itself seems straightforward

Algorithm 1: Computational algorithm for JDHMT model

Input : The tensor $\mathcal{X}$ containing multi-relational triples, the matrix $\mathcal{W}$ containing uni-relational triples.

Output: A (A coupled decomposed embedding matrix preserving both semantics from tensor $\mathcal{X}$ and $\mathcal{W}$), V, C).

/* In the matrix $\mathcal{W}$, row indices are mapped to genes/diseases and coloumn indices are mapped to heterogeneous entities. The cell values are binary for uni-relation present/absent. In the sparse tensor $\mathcal{X}$, rows and coloumns denote genes/disease indices, and the third dimension is mapped to multi-relation type */

Initialize $A, V, \mathcal{C}$.

repeat

  Parameter tuning

  for i *in iteration* **do**

    $A \leftarrow A \cdot \frac{\lambda_{\mathcal{X}} \sum_{k=1}^{K} (\mathcal{X}_k A \mathcal{C}_k^T + \mathcal{X}_k^T A \mathcal{C}_k) + \lambda_W W V^T}{\lambda_{\mathcal{X}} \sum_{k=1}^{K} (A \mathcal{C}_k^T A^T A \mathcal{C}_k + A \mathcal{C}_k A^T A \mathcal{C}_k^T) + \lambda_W A V V^T + \lambda_A A}$ // update rule for A

    $V \leftarrow V \cdot \frac{\lambda_W A^T W}{\lambda_W A^T A V + \lambda_V V}$ // update rule for V

    $\mathcal{C}_k \leftarrow \mathcal{C}_k \cdot \frac{\sum_{k=1}^{K} A^T \mathcal{X}_k A}{\sum_{k=1}^{K} (A^T A \mathcal{C}_k A^T A + \lambda_C \mathcal{C}_k)}$ // update rule for $\mathcal{C}_k$

    $i++$

  end for

until $\mathcal{J}(\theta)$ *convergence*; // here θ denotes the three parameters

return $A, V, \mathcal{C}$

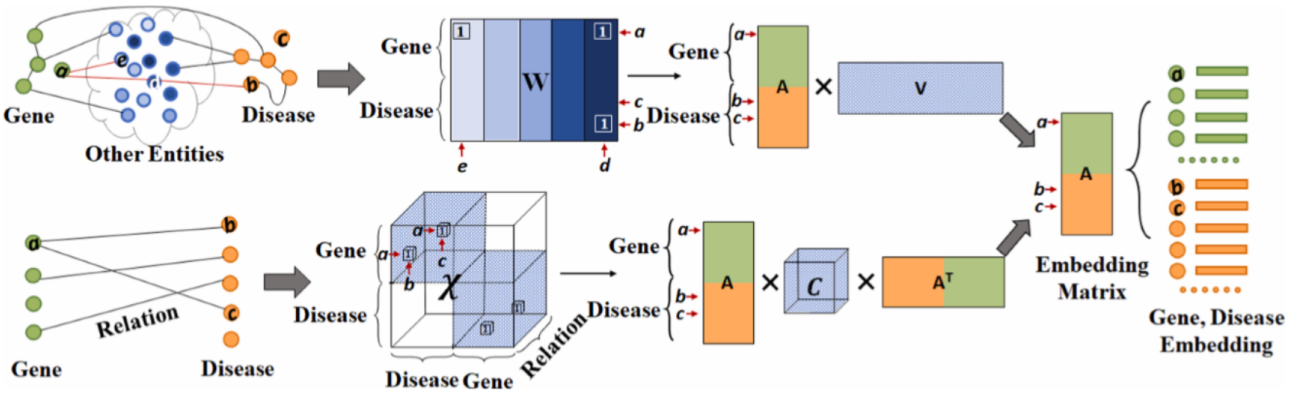

Figure 1: The pictorial representation for Joint Decomposition of Heterogeneous Matrix and Tensor

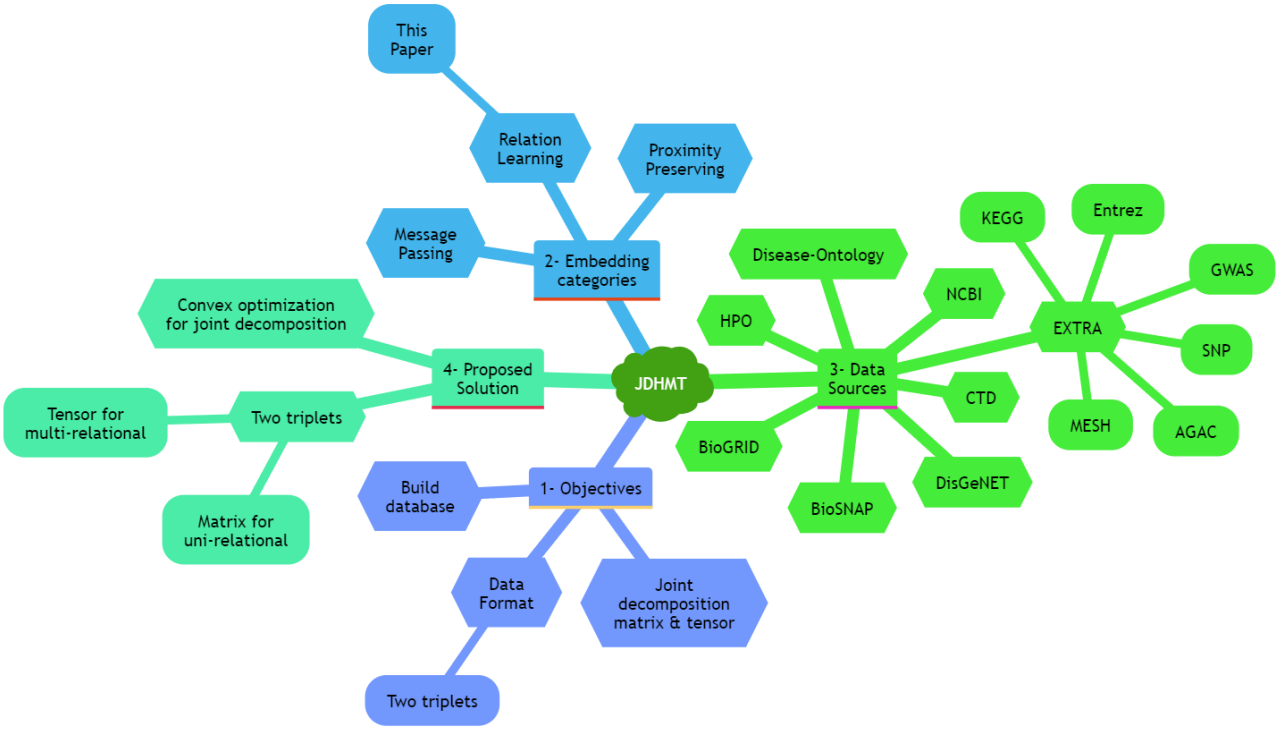

Figure 2: Mindmap

6 Negative Points

- The authors did not perform a compare and contrast exercise for the three paradigms mentioned
- The Embeddings thus generated may not be portable to other applications
- The mapping of gene id or disease id etc. to actual indices of matrix and tensor seems to be critical and potentially non-portable!

7 Questions

1. What is the meaning of concept normalisation for graph node normalisation?
2. The proof of the update rules and the terms A, V, C are yet to be understood.

References

[1] K. Zhou, S. Zhang, Y. Wang, K. B. Cohen, J.-D. Kim, Q. Luo, X. Yao, X. Zhou, and J. Xia, "High-quality gene/disease embedding in a multi-relational heterogeneous graph after a joint matrix/tensor decomposition," *Journal of Biomedical Informatics*, vol. 126, p. 103973, 2022.

[2] J. Xia, "Heterogeneous biological network." Available: `https://hzaubionlp.com/heterogeneous-biological-network/`, Apr 2024. Accessed: 09 July 2024.