Tensor Methods for Multi-layer Graph Embedding

Senior Honors Thesis

submitted by

Anuththari Gamage,

In partial fulfillment of the requirements
for the degree of

Bachelor of Science

in

Electrical Engineering

TUFTS UNIVERSITY

May 2018

ADVISOR: Prof. Shuchin Aeron

Tensor Methods for Multi-layer Graph Embedding

Anuththari Gamage

ADVISOR: Prof. Shuchin Aeron

The abstract goes here.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Graph Embedding

1.1.1 Factorization Methods

1.1.2 Random Walk Methods

1.1.3 Deep Learning

1.2 Multi-layer Graphs

1.3 Multi-layer Graph Embedding

Chapter 2

Review of Related Literature

2.1 Multilayer Graph Embedding

2.1.1 Tensor Factorization methods

2.1.2 Spectral Methods

2.1.3 Other methods

2.2 Pros and Cons of these approaches

Chapter 3

Methodology

3.1 Randomized Shortest Paths

3.1.1 Mathematical Preliminaries

Suppose we are given a weighted, directed, graph $G = \{V, E\}$ where $V = 1, \ldots, n$ defines the set of vertices or nodes, and $E = \{(i, j) \,|\, i \to j\}$ defines edges between nodes. For this graph, we can compute a transition probability matrix $\mathrm{P}^{\mathrm{ref}} = D^{-1}A$. Here, D is the degree matrix containing the degree of each node on the corresponding diagonal entry and A is the adjacency matrix of the graph. Thus, a random walk on the graph will follow a path determined by these transition probabilities.

Consider a particular path on this graph starting at a source node s and a destination node t, denoted by $p_{s\to t}$. Using $\mathrm{P}^{\mathrm{ref}}$, we can compute the probability of this path being taken, $Pr\{p_{s\to t}\}$. If the path taken is $s \to v_1 \to v_2 \to \ldots \to v_m \to t$, then $Pr\{p_{s\to t}\} = \mathrm{P}^{\mathrm{ref}}_{\mathrm{sv_1}}\mathrm{P}^{\mathrm{ref}}_{\mathrm{v_1v_2}} \ldots \mathrm{P}^{\mathrm{ref}}_{\mathrm{v_mt}}$. So, the probability of a path of length m, $p_{s\to t}$ consisting of edges $e_1, e_2, \ldots e_m$ is given by $Pr\{p_{s\to t}\} = \prod_{i=1}^{m} Pr\{e_i\}$.

Since each edge has an associated *cost* determined by the weight on the edge, we can also compute the total cost for a given path $p_{s\to t}$, denoted by $C_{s\to t}$. Suppose the costs of each edge e_i is given by c_{e_i}. This is commonly computed as the reciprocal of the weight on that edge [1]. Then, the total cost of a path $p_{s\to t}$ consisting of the edges $e_1, e_2, \ldots e_m$ is given by $C_{s\to t} = \sum_{i=1}^{m} c_{e_i}$. If we consider only *absorbing* paths,

this cost will be finite. An absorbing path is a path $p_{s\to t}$ such that the destination node t has no outgoing edges except to itself, so a random walk on the path will necessarily terminate.

For notational convenience, we will denote an absorbing path from source node s to destination node t as $\mathcal{P}$, with the probability of the path under the reference probability distribution $\mathrm{P}^{\mathrm{ref}}$ denoted by $\mathrm{P}^{\mathrm{ref}}(\mathcal{P})$ and the cost of traversing the path denoted by $C(\mathcal{P})$. Suppose the set of all such absorbing paths is $\mathcal{P}_{st}$. Then, the *expected cost* of a random walk from a source node s to a destination node t is given by $\sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}^{\mathrm{ref}}(\mathcal{P})C(\mathcal{P})$.

3.1.2 Randomized Shortest Paths

The Randomized Shortest Path (RSP) is defined as the path between two nodes with the minimum expected cost. In order to compute this path, a new probability distribution $\mathrm{P}(\mathcal{P})$ is derived under the following constrained optimization:

$$\mathrm{P}(\mathcal{P}) = \begin{cases} \text{minimize} \sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P})C(\mathcal{P}) \\ \text{subject to} \sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P}) \ln \frac{\mathrm{P}(\mathcal{P})}{\mathrm{P}^{\mathrm{ref}}(\mathcal{P})} = J_0 \end{cases} \tag{3.1}$$

Thus, the new probability distribution is derived by minimizing the expected cost for all possible paths, with the condition that the derived distribution and the reference distribution have a fixed relative entropy J_0 quantified by the Küllback-Leilber divergence between the two distributions. When the parameter J_0 is set to 0, we obtain the original reference probability distribution $\mathrm{P}^{\mathrm{ref}}$. The value of this parameter is determined by a user defined parameter, β, in the actual algorithm.

To obtain a closed-from expression for the RSP probability distribution, we compute the Lagrange function of the above constrained optimization problem:

$$\mathcal{L} = \sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P})C(\mathcal{P}) + \lambda\left[\sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P}) \ln \frac{\mathrm{P}(\mathcal{P})}{\mathrm{P}^{\mathrm{ref}}(\mathcal{P})} - J_0\right] + \mu\left[\sum_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P}) - 1\right] \tag{3.2}$$

For one path $\mathcal{P} \in \mathcal{P}_{st}$, we can solve this as follows:

$$\mathcal{L} = \mathrm{P}(\mathcal{P})C(\mathcal{P}) + \lambda\left[\mathrm{P}(\mathcal{P})\ln\frac{\mathrm{P}(\mathcal{P})}{\mathrm{P}^{\mathrm{ref}}(\mathcal{P})} - J_0\right] + \mu\left[\mathrm{P}(\mathcal{P}) - 1\right]$$

$$\frac{\partial\mathcal{L}}{\partial\mathrm{P}} = C(\mathcal{P}) + \lambda\left[\ln\frac{\mathrm{P}(\mathcal{P})}{\mathrm{P}^{\mathrm{ref}}(\mathcal{P})} + 1\right] + \mu = 0$$

$$\mathrm{P}(\mathcal{P}) = \mathrm{P}^{\mathrm{ref}}(\mathcal{P})e^{-\{\frac{1}{\lambda}(C(\mathcal{P})+\mu)+1\}} = \mathrm{P}^{\mathrm{ref}}(\mathcal{P})e^{-\beta C(\mathcal{P})}$$

We can convert this expression into a probability distribution over all paths by normalizing by the sum of probabilities:

$$\mathrm{P}(\mathcal{P}) = \frac{\mathrm{P}^{\mathrm{ref}}(\mathcal{P})e^{-\beta C(\mathcal{P})}}{\sum\limits_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}^{\mathrm{ref}}(\mathcal{P})e^{-\beta C(\mathcal{P})}} \tag{3.3}$$

3.1.3 Randomized Shortest Path Dissimiliarity Measure

Using the probability distribution for Randomized Shorted Paths derived above, we can define the *dissimiliarity* between two nodes in the following manner. Suppose the expected cost of traversing the randomized shortest path between source node s and destination node t is given by $\overline{C(\mathcal{P}_{st})} = \sum\limits_{\mathcal{P}\in\mathcal{P}_{st}} \mathrm{P}(\mathcal{P})C(\mathcal{P})$. Note that we are using the RSP probability distribution P instead of the reference probability distribution $\mathrm{P}^{\mathrm{ref}}$ here to obtain the minimum expected cost. Then, the *symmetric RSP dissimilarity* between the two nodes s and t is as follows:

$$\Delta_{st}^{\mathrm{RSP}} = \frac{\overline{C(\mathcal{P}_{st})} + \overline{C(\mathcal{P}_{ts})}}{2}$$

The RSP dissimilarity can be thought of as a measure of the distance between two nodes as well, since the expected cost of traversing the randomized shortest path between the nodes increases as the number of edges between them increase, thus driving the RSP dissimilarity to increase. We use this interpretation of $\Delta_{st}^{\mathrm{RSP}}$ to test the efficacy of RSP in graph clustering.

3.1.4 Efficient Computation of the RSP Dissimilarity

Following the derivation of the RSP dissimilarity by Yen et al.[2], an efficient closed-form expression for its computation was derived by Kivimäki et al.[1], which we describe in Algorithm 1. This computation is in done entirely through matrix operations, which lends itself nicely to input graphs in the form of adjacency matrices. The output of the algorithm is the symmetric matrix $\Delta^{\mathrm{RSP}} \in \mathbb{R}^{n \times n}$, in which each entry $\Delta_{ij}^{\mathrm{RSP}}$ gives the RSP dissimilarity between the nodes i and j.

Algorithm 1 RSP Dissimilarity

Input: $\mathrm{P}^{\mathrm{ref}} \in \mathbb{R}^{n \times n}$ (reference transition probability matrix), $C \in \mathbb{R}^{n \times n}$ (cost matrix), β (optimization parameter)
Output: Δ^{RSP} (RSP dissimilarity matrix)
$W = \mathrm{P}^{\mathrm{ref}} \circ e^{-\beta C}$
if $\rho(W) \geq 1$ **then**
  Stop: will not converge
end if
$Z = (I - W)^{-1}$ ($I \in \mathbb{R}^{n \times n}$ is the identity matrix)
$S = (Z[C \circ W]Z) \div Z$ ($\div$ is elementwise division)
$\overline{C} = S - ed_S^T$ ($e \in \mathbb{R}^n$ = all ones vector, $d_S \in \mathbb{R}^n$ = diagonal elements of S)
$\Delta^{\mathrm{RSP}} = (\overline{C} + \overline{C}^T)/2$

3.2 Common Randomized Shortest Paths (C-RSP)

In this work, we extend the core idea behind RSP for multi-layers graphs. Multi-layer or multi-view graphs contain a number of adjacency matrices, which we call layers, defined on the same set of nodes V with different edge distributions on each layer. Thus, if we represent a graph by its vertex and edge sets as $G = \{V, E\}$, then a multi-layer graph is represented as the tensor $\mathcal{G} = \{V, (E_1, \ldots, E_m)\}$ where each layer is given by $G_i = \{V, E_i\}$.

To extend RSP for multil-layer graphs, we derive a common probability distribution, Q, for *all* layers. This is accomplished by once again minimizing the expected cost for all possible paths on all layers, with the condition that the common distribution and the reference probability distribution of each layer, $\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}$ have the same fixed relative entropy. This constrained optimization is represented as follows for a tensor

of m layers, with reference probability distributions $\mathrm{P}_1^{\mathrm{ref}}, \ldots, \mathrm{P}_{\mathrm{m}}^{\mathrm{ref}}$ and cost matrices $C_1, \ldots, C_m$:

$$\mathrm{Q}(\mathcal{P}) = \begin{cases} \text{minimize} \sum_{i=1}^{m} \sum_{\mathcal{P} \in \mathcal{P}_{st}} \mathrm{Q}(\mathcal{P}) C_i(\mathcal{P}) \\ \text{subject to} \sum_{i=1}^{m} \sum_{\mathcal{P} \in \mathcal{P}_{st}} \mathrm{Q}(\mathcal{P}) \ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}(\mathcal{P})} = J_0 \end{cases} \tag{3.4}$$

Consider a multi-layer graph $\mathcal{G}$ with $m = 2$ layers, $G_1 = \{V, E_1\}$ and $G_2 = \{V, E_2\}$, with the reference transition probability distributions $\mathrm{P}_1^{\mathrm{ref}}, \mathrm{P}_2^{\mathrm{ref}}$ and cost matrices C_1, C_2 respectively. To derive the common distribution $\mathrm{Q}(\mathcal{P})$, we use the Lagrange function as follows:

$$\mathcal{L} = \sum_{i=1}^{m} \sum_{\mathcal{P} \in \mathcal{P}_{st}} \mathrm{Q}(\mathcal{P}) C_i(\mathcal{P}) + \lambda \left[\sum_{\mathcal{P} \in \mathcal{P}_{st}} \mathrm{Q}(\mathcal{P}) \ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}(\mathcal{P})} - J_0 \right] + \mu \left[\sum_{\mathcal{P} \in \mathcal{P}_{st}} \mathrm{Q}(\mathcal{P}) - 1 \right] \tag{3.5}$$

Considering only one path, we obtain the following:

$$\mathcal{L} = \mathrm{Q}(\mathcal{P}) C_1(\mathcal{P}) + \mathrm{Q}(\mathcal{P}) C_2(\mathcal{P}) + \lambda \left[\mathrm{Q}(\mathcal{P}) \ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_1^{\mathrm{ref}}(\mathcal{P})} - J_0 \right] + \lambda \left[\mathrm{Q}(\mathcal{P}) \ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_2^{\mathrm{ref}}(\mathcal{P})} - J_0 \right] + \mu \left[\mathrm{Q}(\mathcal{P}) - 1 \right]$$

$$\frac{\partial \mathcal{L}}{\partial \mathrm{Q}} = C_1(\mathcal{P}) + C_2(\mathcal{P}) + \lambda \left[\ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_1^{\mathrm{ref}}(\mathcal{P})} + \ln \frac{\mathrm{Q}(\mathcal{P})}{\mathrm{P}_2^{\mathrm{ref}}(\mathcal{P})} + 2 \right] + \mu = 0$$

$$\ln \left[\frac{Q^2(\mathcal{P})}{\mathrm{P}_1^{\mathrm{ref}}(\mathcal{P}) \mathrm{P}_2^{\mathrm{ref}}(\mathcal{P})} \right] = -\frac{1}{\lambda} \left[(C_1 + C_2) + 2 \right] - \mu$$

$$\mathrm{Q}(\mathcal{P}) = \sqrt{\mathrm{P}_1^{\mathrm{ref}}(\mathcal{P}) \mathrm{P}_2^{\mathrm{ref}}(\mathcal{P})} \left[e^{-\frac{1}{2} \{ \frac{1}{\lambda} [(C_1 + C_2) + 2] - \mu \}} \right]$$

$$\mathrm{Q}(\mathcal{P}) = \sqrt{\mathrm{P}_1^{\mathrm{ref}}(\mathcal{P}) \mathrm{P}_2^{\mathrm{ref}}(\mathcal{P})} \left[e^{-\beta (C_1 + C_2)} \right]$$

Extending this derivation to m layers and normalizing as before, we obtain the following expression for the C-RSP probability distribution for a single path:

$$\mathrm{Q}(\mathcal{P}) = \frac{\sqrt[m]{\Pi_{i=1}^{m} \mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}(\mathcal{P})} e^{-\beta \left(\sum_{i=1}^{m} C_i \right)}}{\sum_{\mathcal{P} \in \mathcal{P}_{st}} \sqrt[m]{\Pi_{i=1}^{m} \mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}(\mathcal{P})} e^{-\beta \left(\sum_{i=1}^{m} C_i \right)}} \tag{3.6}$$

3.2.1 C-RSP Dissimiliarity Measure

Using the derived common probablility distribution, $\mathrm{Q}(\mathcal{P})$, we can compute a dissimilarity measure $\Delta^{\mathrm{C-RSP}}$ for multi-layer graphs following an approach similar to that detailed in Section 3.1.4 above. Note that we can use the same algorithm used for computing RSP if we were to have a single reference probability matrix $\mathrm{P}^{\mathrm{ref}}$ and a single cost matrix C instead of the tensors associated with a multi-layer graph. Using the expression for $\mathrm{Q}(\mathcal{P})$ derived in Equation 3.6, we can *combine* the individual layers of these tensors to obtain these matrices as detailed below.

Let $\mathbf{P}$ denote the combined reference transition probability matrix and $\mathbf{C}$ denote the combined cost matrix. By comparing Equations 3.3 and 3.6, we obtain,

$$\mathbf{P} = \sqrt[m]{\prod_{i=1}^{m} \mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}(\mathcal{P})} \tag{3.7}$$

$$\mathbf{C} = \sum_{i=1}^{m} C_i \tag{3.8}$$

Note that in Equation 3.6, we derive the probability of an individual path and not the entire set of possible paths. Thus, we need to take care to omit instances when the path does not exists given a particular $\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}$, which occurs when an entry in any $\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}$ is zero. Also note that the multiplication in this expression is an *elementwise* multiplication across the layers of the reference probability tensor rather than a matrix multiplication of the layers. Furthermore, the m^{th} root is also taken elementwise.

To account for all these subtleties in the computation of $\mathbf{P}$, we can better represent this matrix in the following manner:

$$\mathbf{P}_{ij} = \begin{cases} \sqrt[M]{\prod_{k=1}^{M} (\mathrm{P}_{\mathrm{k}}^{\mathrm{ref}})_{ij}} & \exists (\mathrm{P}_{\mathrm{k}}^{\mathrm{ref}})_{ij} \neq 0, \text{ where } M = |\{\mathrm{P}_{\mathrm{k}}^{\mathrm{ref}})_{ij} \neq 0\}| \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

This manner of combining the different $\mathrm{P}_{\mathrm{i}}^{\mathrm{ref}}$ matrices does not guarantee a row-stochastic matrix, which is necessary for it to be a probability distribution. Thus,

the resulting matrix $\mathbf{P}$ must be further manipulated to obtain a row-stochastic matrix. This can be achieved easily by successive division of each row and column of the matrix by their respective row and column sums until convergence.

Using these steps, we obtain a combined reference probability matrix $\mathbf{P}$ and a combined cost matrix $\mathbf{C}$ that can then be used in the original RSP algorithm to obtain the C-RSP dissimilarity measure $\Delta^{\mathrm{C-RSP}}$, as detailed in Algorithm 2 below.

Algorithm 2 C-RSP Dissimilarity

Input: $\{\mathrm{P}_1^{\mathrm{ref}}, \ldots, \mathrm{P}_{\mathrm{m}}^{\mathrm{ref}}\} \in \mathbb{R}^{n \times n \times m}$ (reference transition probability tensor), $\{C_1, \ldots, C_m\} \in \mathbb{R}^{n \times n \times m}$ (cost tensor), β (optimization parameter)
Output: Δ^{RSP} (C-RSP dissimilarity matrix)
  $\mathbf{P} = \{\mathrm{P}_1^{\mathrm{ref}}, \ldots, \mathrm{P}_{\mathrm{m}}^{\mathrm{ref}}\}$ **combined** as given in Equation 3.9
  while $\mathbf{P}$ not row-stochastic **do**
    Divide each row by the row sum
    Divide each column by the column sum
  end while
  $\mathbf{P} = \mathrm{stochastize}(\mathbf{P})$
  $\mathbf{C} = \sum_{i=1}^{m} C_i$
  $W = \mathrm{P}^{\mathrm{ref}} \circ e^{-\beta C}$
  if $\rho(W) \geq 1$ **then**
    Stop: will not converge
  end if
  $Z = (I - W)^{-1}$ ($I \in \mathbb{R}^{n \times n}$ is the identity matrix)
  $S = (Z[C \circ W]Z) \div Z$ ($\div$ is elementwise division)
  $\overline{C} = S - e d_S^T$ ($e \in \mathbb{R}^n$ = all ones vector, $d_S \in \mathbb{R}^n$ = diagonal elements of S)
  $\Delta^{\mathrm{C-RSP}} = (\overline{C} + \overline{C}^T)/2$

Chapter 4

Evaluation and Findings

Chapter 5

Conclusion

And future work

[2]

Bibliography

[1] Ilkka Kivimäki, Masashi Shimbo, and Marco Saerens. "Developments in the theory of randomized shortest paths with a comparison of graph node distances". In: *Physica A: Statistical Mechanics and its Applications* 393 (2014), pp. 600–616. ISSN: 03784371. DOI: 10.1016/j.physa.2013.09.016. arXiv: 1212.1666. URL: http://dx.doi.org/10.1016/j.physa.2013.09.016.

[2] Luh Yen et al. "A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (2008), p. 785. DOI: 10.1145/1401890.1401984. URL: http://dl.acm.org/citation.cfm?doid=1401890.1401984.