Proximity Preservation and Crossing-Minimization for Graph Embedding

Amina Shabbeer[1] and Kristin P. Bennett[1,2]

[1] Computer Science Department, Rensselaer Polytechnic Institute
[2] Mathematical Sciences Department, Rensselaer Polytechnic Institute

We propose a novel approach to embedding heterogeneous data in high-dimensional space characterized by a graph. Targeted towards data visualization, the objectives of the embedding are two-fold: (i) preserve proximity relations as measured by some embedding objective, and (ii) simultaneously optimize an aesthetic criterion, no edge-crossings in the embedding, to create a clear representation of the underlying graph structure. This method is applicable for graphs where the nodes represent objects that have their own intrinsic properties with associated distances or similarity measures that describe implicit relations between all pairs of nodes. The motivating application for this work was generating visualizations for phylogenetic trees of strains of *Mycobacterium tuberculosis* with defined genetic distances between every pair of strains. It is often desirable, that drawings of such graphs map nodes from high-dimensional feature space to low-dimensional vectors that preserve these pairwise distances. This desired quality is frequently expressed as a function of the embedding and then optimized, e.g. in Multidimensional Scaling (MDS), the goal is to minimize the difference between the actual pairwise distances and Euclidean distances in the embedding for all nodes. However, layouts that preserve proximity relations can have a large number of edge-crossings obfuscating the relationships between nodes making the graph difficult to understand and interpret. It is therefore desirable to minimize edge crossings. This is a challenging problem in itself; determining the minimum number of crossings for a graph is NP-complete [2].

The principle contributions of this paper are (i) expressing edge-crossing minimization as a *continuous* optimization problem (ii) An iterative penalty algorithm that elegantly incorporates the nonconvex nonsmooth constraints arising from the edge-crossing minimization formulation into an optimization routine for embedding objectives e.g. stress majorization in MDS [1].

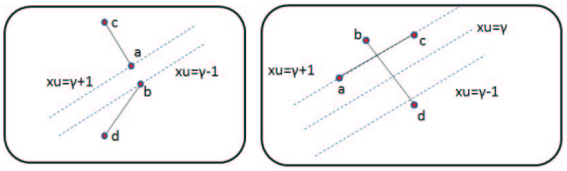

Fig. 1: In (a) Edge $\mathcal{A}$ from a to c and edge $\mathcal{B}$ from b to d do not cross. Any line between $xu - \gamma = 1$ and $xu - \gamma = -1$ strictly separates the edges. Using a soft margin, the plane in (b) $xu - \gamma = 0$ separates the plane into half spaces that should contain each edge.

The key theoretical insight of the paper is that the condition that two edges do **not** cross is equivalent to the feasibility of a system of nonlinear inequalities. Two edges do not intersect iff the following system of equations has **no solution**:

$$\nexists \quad \delta_A,\, \delta_B \text{ s.t.} \quad A'\delta_A = B'\delta_B \quad e'\delta_A = 1 \quad e'\delta_B = 1 \quad \delta_A \geq 0 \quad \delta_B \geq 0 \tag{1}$$

where e is a vector of ones of appropriate dimension and $A = \begin{bmatrix} a_x & a_y \\ c_x & c_y \end{bmatrix}$ and $B = \begin{bmatrix} b_x & b_y \\ d_x & d_y \end{bmatrix}$. We prove this using a theorem of the alternative: Farkas' lemma [3]. Therefore, two edges **do not** intersect iff $||(-Au + (1+\gamma)e)_+|| + ||(Bu + (1-\gamma)e)_+|| = 0$ where $(z)_+ = max(0, z)$. Geometrically the theorem states that two edges (or more generally two polyhedrons) do not intersect if and only if there exists a hyperplane that strictly separates the extreme points of $\mathcal{A}$ and $\mathcal{B}$. Figure 1 illustrates that when this system is satisfied, any plane that lies between $xu+\gamma = 1$ and $xu+\gamma = -1$ strictly separates the two edges, and the edges do not intersect. This formulation bears resemblance to the parallel hyperplanes used to find maximum margin hyperplanes

in SVM [4]. The no-edge-crossing constraint corresponds to introducing a hyperplane and requiring each edge to lie in opposite half spaces. The constraints can be generalized to remove intersections of general convex polygons including node-edge and node-node intersections. The proposed edge-crossing constraints and iterative penalty algorithm can be readily adapted to other supervised and unsupervised optimization-based embedding or dimensionality reduction methods.

While edge crossing minimization can be utilized in conjunction with any optimization-based embedding objective, here we demonstrate the approach on multidimensional scaling by modifying the stress majorization algorithm to include penalties for edge crossings. An alternating iterative penalty algorithm, Alternating Majorization Algorithm (AMA) is developed, to minimize stress subject to a large number of non-convex non-smooth constraints. The algorithm is applied to a problem in tuberculosis molecular epidemiology, creating 'spoligoforests' for visualizing genetic relatedness between strains characterized by fifty-five biomarkers with associated non-Euclidean genetic distances of the *Mycobacterium tuberculosis* complex as shown in Fig. 2. Comparisons with other dimensionality reduction techniques and classical graph drawing algorithms are made to demonstrate the efficacy of the method. The performance of the algorithm is also demonstrated on a challenging test suite of randomly generated graphs. Computational results demonstrate that this approach is practical and tractable. Animations of the algorithm illustrating how the edge crossing penalty progressively transform the graphs are provided *http://www.cs.rpi.edu/∼shabba/FinalGD/*. [3]

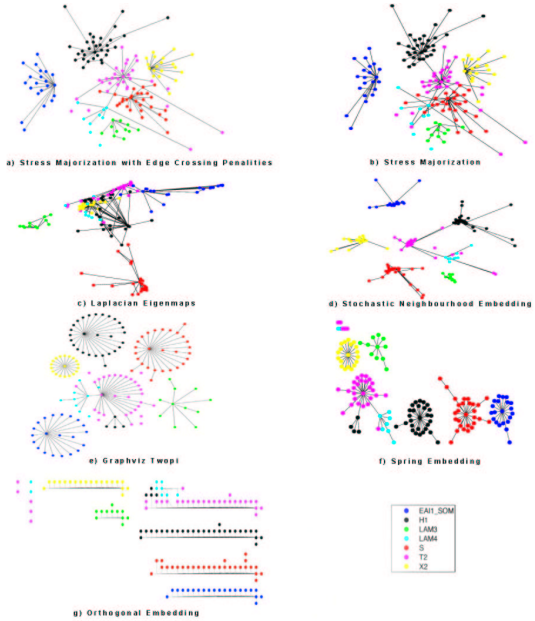

Fig. 2: Embeddings of spoligoforests of SpolDB4 sublineages by 7 algorithms. (e-h) generated by classical graph drawing algorithms fail to represent genetic distances. Embeddings generated by dimensionality reduction techniques (b-d) have a large number of crossings, moreover all pairwise distances are not preserved in (c-d). Graph (b), that optimizes the MDS objective and generated using Neato, preserves proximity relations but has edge-crossings. In graph (a), the proposed approach eliminates all edge crossings with little change in the overall stress. Note how in graph (a), the radial structure emerges naturally when both distances and the graph structure are considered.

References

1. J. De Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of classification*, 5(2):163–180, 1988.
2. M.R. Garey and D.S. Johnson. Crossing number is np-complete. *SIAM Journal on Algebraic and Discrete Methods*, 4:312, 1983.
3. O.L. Mangasarian. *Nonlinear programming*. Society for Industrial Mathematics, 1994.
4. V.N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[3] Preference:Oral/Poster