1 Rebuttal to Question 7 of the reviwer q3r3 Part 2

In section 4.1, we state that: In Appendix B1, we also give practical evidence of the need to solve Trace-Ratio problems in an SGD context, instead of solving the original Trace problem in Eq.2. We also justify the convenience of conditioning $\mathbf{U}$ to $\mathbf{A}$, $\mathbf{U} = f_\theta(\mathbf{A})$. Actually, this setting is inspired by how the LINKX method exploits the graph topology.

We clarify this point with the following discussion:
Parameterizing U. One of the key elements of our method is that $\mathbf{U} = f_\Theta(\mathbf{A})$, where f_Θ is an MLP whose output non-linear units γ are of the type Tanh. For simplicity, let $\mathbf{U} = \gamma(\mathbf{AV})$, where $\mathbf{V}$ has dimension $n \times p$. Then, following LINKX we have:

$$(\mathbf{AV})_{ip} = \mathbf{A}_{i:}\mathbf{V}_{:p} = \sum_{j\in\mathcal{N}_i} \mathbf{V}_{jp} \ . \tag{1}$$

For a better understanding of the role of $\mathbf{AV}$, we observe that solving the Trace-Ratio problem is equivalent to solving the minimization problem

$$\mathcal{L}'_D = \frac{Tr[\mathbf{U}^T\triangle\mathbf{U}]}{Tr[\mathbf{U}^T\mathbf{D}\mathbf{U}]} \ , \ \nabla_{\mathbf{U}}(\mathcal{L}'_D) = \frac{2\left(\triangle - \rho\mathbf{D}\right)\mathbf{U}}{Tr[\mathbf{U}^T\mathbf{D}\mathbf{U}]} \ , \tag{2}$$

s.t. $\mathbf{U}\mathbf{U}^T = \mathbf{I}$, where $\triangle = \mathbf{D} - \mathbf{A}$ is the graph Laplacian. Expanding the numerator of $\mathcal{L}'_D$, we have

$$Tr[\mathbf{U}^T\triangle\mathbf{U}] = \sum_{p=1}^{P} \mathbf{u}_p^T\triangle\mathbf{u}_p = \sum_{p=1}^{P}\sum_{i\sim j}\left(\mathbf{u}_{ip} - \mathbf{u}_{jp}\right)^2 ; . \tag{3}$$

Then, plugging $\mathbf{AV}$ in the trace, we obtain:

$$\mathbf{u}_p^T\triangle\mathbf{u}_p = \sum_{i\sim j}(\gamma(\sum_{j\in\mathcal{N}_i}\mathbf{V}_{jp}) - \gamma(\sum_{k\in\mathcal{N}_j}\mathbf{V}_{kp}))^2; . \tag{4}$$

As a result, minimizing $\mathcal{L}'_D$ results in learning MLP weights $\mathbf{V}$ enforcing that the two nodes of an edge (i, j) *have common neighbors*. This means that our energy minimization leads to infer intra-class edges. As a result, the role of the parameterization is to enforce the cohesiveness of the cluster. This is consistent with the fact that we achieve better results with this parameterization.