Appendix of "Clustering Ensemble via Structured Hypergraph Learning"

Peng Zhou[a,b], Xia Wang[a], Liang Du[c], Xuejun Li[a]

[a]*School of Computer Science and Technology, Anhui University, Hefei 230601, China.*
[b]*The State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China*
[c]*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*

Proof of Theorem 1

Theorem 1. *Given any hypergraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$ with n nodes, if the rank of its Laplacian matrix $\mathbf{L}$, which is defined as $\mathbf{L} = \mathbf{I} - \mathbf{D}_v^{-\frac{1}{2}}\mathbf{Y}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{Y}^T\mathbf{D}_v^{-\frac{1}{2}}$, is $n - c$, then $\mathcal{G}$ contains exact c connective components.*

Proof. Before proving this Theorem, we provide the following lemma:

Lemma 1. *Given any connective hypergraph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{W}\}$ with n nodes (i.e., $\mathcal{G}$ only contains one connective component), the rank of its Laplacian matrix is $n - 1$.*

Proof. Denote $\mathbf{H}$ as the incidence matrix of $\mathcal{G}$, we can compute its Laplacian matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}}$. For any vector $\mathbf{x} \in \mathbb{R}^n$, we have

$$
\begin{aligned}
\mathbf{x}^T\mathbf{L}\mathbf{x} &= \mathbf{x}^T(\mathbf{I} - \mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}})\mathbf{x} = \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}}\mathbf{x} \\
&= \sum_{u\in\mathcal{V}} \mathbf{x}(u)^2 \sum_{e\in\mathcal{E}} \frac{\mathbf{W}(e)\mathbf{H}(u,e)}{\mathbf{D}_v(u)} \sum_{v\in\mathcal{V}} \frac{\mathbf{H}(v,e)}{\mathbf{D}_e(e)} - \sum_{e\in\mathcal{E}} \sum_{u,v\in\mathcal{V}} \frac{\mathbf{x}(u)\mathbf{H}(u,e)\mathbf{W}(e)\mathbf{H}(v,e)\mathbf{x}(v)}{\sqrt{\mathbf{D}_v(u)\mathbf{D}_v v}\mathbf{D}_e(e)} \\
&= \frac{1}{2}\sum_{e\in\mathcal{E}} \sum_{u,v\in\mathcal{V}} \frac{\mathbf{W}(e)\mathbf{H}(u,e)\mathbf{H}(v,e)}{\mathbf{D}_e(e)} \left(\frac{\mathbf{x}(u)}{\sqrt{\mathbf{D}_v(u)}} - \frac{\mathbf{x}(v)}{\sqrt{\mathbf{D}_v(v)}} \right)^2 \qquad (1)
\end{aligned}
$$

Email addresses: zhoupeng@ahu.edu.cn (Peng Zhou), e19201043@stu.ahu.edu.cn (Xia Wang), csliangdu@gmail.com (Liang Du), xjli@ahu.edu.cn (Xuejun Li)

where the third equation is due to the definition of $\mathbf{D}_v$ and $\mathbf{D}_e$ and the fourth equation is due to the completing square formula.

Obviously, for any node u, if $\mathbf{x}(u) = \sqrt{\mathbf{D}_v(u)}$, we have $\mathbf{x}^T\mathbf{L}\mathbf{x} = 0$. Therefore, this $\mathbf{x}$ is an eigenvector of $\mathbf{L}$ whose corresponding eigenvalue is 0. Since we aim to prove that $rank(\mathbf{L}) = n - 1$, we need to prove that $\mathbf{L}$ does not have any other eigenvector $\mathbf{x}'$ (which is linearly independent with $\mathbf{x}$), whose corresponding eigenvalue is also 0.

We use the proof by contradiction. We assume that there exists such $\mathbf{x}'$ whose corresponding eigenvalue is also 0, and $\mathbf{x}'$ is linearly independent with $\mathbf{x}$, i.e., there does not exist a constant scalar t such that $\mathbf{x}' = t\mathbf{x}$. Since $\mathbf{x}'$'s corresponding eigenvalue is 0, we have

$$\mathbf{x}'^T\mathbf{L}\mathbf{x}' = \frac{1}{2}\sum_{e\in\mathcal{E}}\sum_{u,v\in\mathcal{V}}\frac{\mathbf{W}(e)\mathbf{H}(u,e)\mathbf{H}(v,e)}{\mathbf{D}_e(e)}\left(\frac{\mathbf{x}'(u)}{\sqrt{\mathbf{D}_v(u)}} - \frac{\mathbf{x}'(v)}{\sqrt{\mathbf{D}_v(v)}}\right)^2 = 0, \quad (2)$$

which means for any two nodes u and v, if there exists a hyperedge e such that $u \in e$ and $v \in e$ (i.e., $\mathbf{H}(u,e) = \mathbf{H}(v,e) = 1$), then $\frac{\mathbf{x}'(u)}{\sqrt{\mathbf{D}_v(u)}} = \frac{\mathbf{x}'(v)}{\sqrt{\mathbf{D}_v(v)}}$.

Since $\mathcal{G}$ is a connective hypergraph, which means for any two nodes u and v, there exists at least one path $e_1, \cdots, e_r$ such that $u \in e_1$, $u_1 \in e_1$, $u_1 \in e_2$, $u_2 \in e_2$, $u_2 \in e_3$, ..., $u_r \in e_r$, and $v \in e_r$. Then, we have $\frac{\mathbf{x}'(u)}{\sqrt{\mathbf{D}_v(u)}} = \frac{\mathbf{x}'(u_1)}{\sqrt{\mathbf{D}_v(u_1)}} = \cdots = \frac{\mathbf{x}'(v)}{\sqrt{\mathbf{D}_v(v)}}$. Therefore, $\mathbf{x}' = t\sqrt{diag(\mathbf{D}_v)} = t\mathbf{x}$, which is contradict with the assumption that $\mathbf{x}'$ is linearly independent with $\mathbf{x}$. This concludes the proof of Lemma 1. $\square$

Now come back to the proof of Theorem 1. Suppose $\mathcal{G}$ has c' connective components $\mathcal{G}_1, \cdots, \mathcal{G}_{c'}$. It is easy to verify that the incidence matrix $\mathbf{Y}$ of $\mathcal{G}$ can be written as the direct sum of incidence matrices $\mathbf{Y}_1, \cdots, \mathbf{Y}_{c'}$ of $\mathcal{G}_1, \cdots, \mathcal{G}_{c'}$, i.e., $\mathbf{Y} = \mathbf{Y}_1 \oplus \cdots \oplus \mathbf{Y}_{c'}$ where $\oplus$ denotes the direct sum. Then the Laplacian matrix $\mathbf{L}$ can also be written as the direct sum of $\mathbf{L}_1, \cdots, \mathbf{L}_{c'}$. Therefore, we have

$$rank(\mathbf{L}) = rank(\mathbf{L}_1 \oplus \cdots \oplus \mathbf{L}_{c'}) = \sum_{p=1}^{c'} rank(\mathbf{L}_p). \quad (3)$$

According to Lemma 1, for $\mathcal{G}_p$ which has n_p instances, we have $rank(\mathbf{L}_p) = n_p - 1$.

Then

$$rank(\mathbf{L}) = \sum_{p=1}^{c'} rank(\mathbf{L}_p) = \sum_{p=1}^{c'} n_p - c' = n - c'. \tag{4}$$

Note that $rank(\mathbf{L}) = n - c$, we have $c' = c$, i.e., $\mathcal{G}$ has c connective components, which concludes the proof of Theorem 1. $\square$